Análise de Fiabilidade das Ordens de Trabalho

Sérgio Gabriel Pontes de Jesus

Dissertação de Mestrado

Orientador na FEUP: Dr. Armando Luís Ferreira Leitão

Orientador na Empresa: Eng. Eduardo Miguel Lage Dixo de Sousa

Mestrado Integrado em Engenharia Mecânica

2018-07-02

‘‘Escolhe um trabalho que gostes e não terás
que trabalhar nem um dia na tua vida’’
Confúcio

Resumo

A importância da manutenção e gestão de ativos têm vindo a crescer no que diz respeito ao sucesso de uma empresa na indústria, sendo uma fonte de custos de atividade que pode ser facilmente reduzida, mas que muitas vezes é esquecida em favor de outros aspetos da gestão da produção. As ordens de trabalho são uma unidade fundamental para o estudo da fiabilidade de equipamentos, pois fornecem informações sobre o modo de avaria bem como os respetivos tempos de avaria. No entanto, a análise destes documentos ainda é feita de modo manual, e, muitas vezes situacional, não tirando proveito de toda a informação contida nas bases de dados das ordens de trabalho e alocando recursos para esta tarefa.

Com este projeto pretende-se estudar soluções e desenvolver um protótipo para a análise de ordens de trabalho de forma automatizada. Para atingir este objetivo foram desenvolvidas soluções de *machine learning* aplicadas ao processamento de linguagem natural, para tarefas como a classificação do tipo de ordem de trabalho e modo de falha de um equipamento. A partir dos dados processados, foi feita uma análise fiabilística, assumindo independência entre as falhas, de modo a obter a caracterização da taxa de avarias do equipamento.

Para o estudo foram utilizados três conjuntos de dados provenientes de diferentes fontes; nos dois primeiros casos, estão explícitas as ordens de trabalho realizadas para um dado equipamento e durante um dado período de tempo de todos os equipamentos respetivamente, onde foram aplicadas as soluções estudadas para tratamento de dados e análise de linguagem natural. No terceiro conjunto estão explícitos os dados já tratados para uma análise fiabilística tradicional.

Na realização da dissertação foi possível concluir o potencial da utilização de modelos de *machine learning* aplicados a quantidades de dados a uma escala industrial (*Big Data*), para a extração de informação relevante, apresentando soluções viáveis face às metodologias tradicionais.

Palavras-Chave: Modelos de Previsão, *Machine Learning*, Processamento de Linguagem Natural, Análise de Fiabilidade, Ordens de Trabalho

Reliability Analysis through Work Orders

Abstract

The importance of maintenance and asset management has been growing when it comes to the success of a company in the industry, being a source of activity costs that can be easily reduced, but many times forgoten in favour of other areas of production management. Work orders are one of the fundamental pieces for reliabilty studies in production equipments, as they give us information about failure modes, as well as their respective times. However, the analysis of these documents is still done manually, and many time situational, not benefiting from all the information contained in the work orders' database and alocating unecessary resources to this task.

The scope of this project is to study solutions and develop a prototype for automated work order analysis. To achieve this, machine learning solutions applied to Natural Language Processing (NLP) have been developed, to classify the work order type and the equipment's failure mode. After the data processing task, a reliability analysis has been carried, assuming independence between failure modes, to obtain the hazard rate of the equipment.

For this study three datasets from different sources were used; In the first two cases, the work orders are from an equipment and from a time period of several equipments, respectively, where the studied solutions for data pre-processing and natural language processing were used. For the latter dataset, the data was previously treated, and traditional reliability and survival analysis methods were applied.

During this project it was possible to learn the applicability of machine learning models in industrial size datasets (Big Data), to extract relevant information from data, showing to be a viable alternative to current analysis solutions.

Key Words: Prediction Models, Machine Learning, Natural Language Processing, Reliability Analysis, Work Orders

Agradecimentos

A todas as pessoas na EQS que me ajudaram durante a realização da dissertação em ambiente empresarial, pelo apoio, motivação e esclarecimento de qualquer duvída que surgisse, nas mais diversas áreas.

Ao Eng. Eduardo Dixo, orientador na empresa, por toda a disponibilidade que teve para me acompanhar, em momentos mais e menos sérios, e por se revelar um verdadeiro amigo.

Ao Dr. Armando Leitão pela importante ligação ao meio académico, boas práticas, conselhos e revisão na realização deste projeto.

À minha familia, Mãe, Pai e Ana, por todos os bons momentos e carinho durante a minha vida.

Índice de Conteúdos

Siglas

ANN – *Artificial Neural Network*

AUC – *Area under the curve*

CBTW – *Category Based Term Weights*

FN – *False Negative*

FP– *False Positive*

MLE – *Maximum Likelihood Estimation*

MTBF – *Mean Time Between Failure*

MTTF – *Mean Time to Failure*

NB - *Naïve Bayes*

OT – Ordem de Trabalho

ReLU – *Rectified Linear Unit*

ROC – *Receiver Operating Characteristic*

SVM – *Support Vector Machine*

TF-IDF – *Term Frequency-Inverse Document Frequency*

TN – *True Negative*

TP – *True Positive*

Índice de Figuras

Índice de Tabelas

1 Introdução

Neste primeiro capítulo, será feita uma introdução ao projeto de dissertação "Análise de Fiabilidade das Ordens de Trabalho" realizado na empresa EQS - Engenharia, Qualidade e Segurança, Lda., no âmbito da unidade curricular Dissertação pertencente ao plano de estudos do Mestrado Integrado em Engenharia Mecânica, opção de Gestão da Produção da Faculdade de Engenharia da Universidade do Porto.

Será apresentada a motivação que levou o desenvolvimento do projeto e em que disciplinas este se enquadra, uma breve descrição das entidades que ofereceram a oportunidade desta Dissertação em ambiente empresarial, as principais etapas e objetivos que foram traçados para o projeto e a consequente metodologia usada para os alcançar, e finalmente será feita uma descrição de como a dissertação se apresentará estruturada.

1.1 Enquadramento do Projeto e Motivação

No presente, a quantidade, variedade e complexidade de equipamentos numa dada unidade de produção tem uma tendência crescente, e, com o desenvolvimento tecnológico, a informação é mais detalhada e abundante. Com o aumento da informação é também necessária a melhoria dos processos de análise e tratamento dos dados existentes, de modo a ser obtida informação objetiva, pertinente e correta do estado e o comportamento ao longo do tempo desses mesmos equipamentos, de uma forma sistematizada e em tempo útil.

Durante as operações de manutenção são realizadas ordens de trabalho, que são documentos datados onde são descritos o estado e as operações realizadas num dado equipamento ou conjunto de equipamentos. Estes documentos são providos de espaços destinados à escrita em linguagem natural (linguagem não normalizada) para o preenchimento do trabalhador. Estes documentos são posteriormente transcritos para bases de dados, e analisados manualmente por trabalhadores especializados. Quando realizada, esta análise é um processo monótono, morado e de custo elevado, que pode ser otimizado e automatizado com ferramentas adequadas.

Este projeto enquadra-se no âmbito de várias disciplinas de Mecânica e de Gestão de Produção, sendo mais aprofundado nas de Gestão de Manutenção, Gestão de Qualidade Total, e Programação, bem como Estatística e Álgebra.

1.2 Apresentação da empresa EQS e do projeto UNO

1.2.1 A EQS - Engenharia, Qualidade e Segurança, Lda.

A EQS - Engenharia, Qualidade e Segurança, Lda. é uma empresa constituída em 2005, sediada na Maia, que presta serviços de engenharia nas áreas da inspeção, ensaios e certificação de ativos, formação e consultoria em várias áreas da gestão (Qualidade, ativos, sistemas de gestão, gestão de *software*, entre outros) e desenvolvimento de soluções tecnológicas e *software*. A EQS pretende-se diferenciar através da inovação e utilização dos processos mais avançados tecnologicamente

Os serviços são prestados a uma variedade de indústrias, desde a indústria transformadora à de telecomunicações, passando também pela automóvel e construção. O maior consumidor é a indústria energética, em particular a do petróleo e petroquímica.

A empresa apresenta mercado nacional e internacional, sendo criados desde a fundação novos postos em Lisboa, Sines e Luanda.

Figura 1 - Logótipo da EQS

1.2.2 O Projeto UNO

O projeto UNO – *Means you know* surgiu da necessidade de atualizar os sistemas tradicionais de gestão de ativos, para um modelo mais inovador, em forma de *software* como serviço (SaaS), onde é concentrada a infraestrutura e recursos necessários para o funcionamento da plataforma *Web*. Esta é uma forma de criar um serviço mais dinâmico e adaptado ao cliente, onde as soluções serão moldadas aos requisitos e especificações pretendidos.

O *software* abrange não só o armazenamento e gestão da informação em *cloud*, como também processamento e análise (*Analytics*), recorrendo a métodos de Inteligência Artificial para esse efeito.

O projeto UNO pertence e decorre nas instalações da EQS e foi um dos vencedores do financiamento do programa europeu de inovação *Horizon 2020*.

Figura 2 - Logótipo do projeto UNO

1.3 Objetivos do projeto

Esta dissertação tem como objetivo final a pesquisa e desenvolvimento de algoritmos programáveis e aplicáveis ao projeto UNO para a análise fiabilística de equipamentos a partir de dados obtidos em ordens de trabalho. Este objetivo principal pode ser partido em duas fases, sendo a primeira o tratamento dos dados, e a segunda a análise fiabilística.

Os marcos de cada uma das fases são, respetivamente, os dados tratados e os resultados fiabilísticos. Um objetivo secundário deste projeto é o estudo e a quantificação do erro associado a cada uma das operações e naturalmente a sua redução.

1.4 Metodologia

A metodologia durante o projeto foi estruturada em quatro etapas diferentes.

Numa fase inicial, fez-se uma introdução de várias ferramentas e metodologias que são utilizadas pela empresa no projeto, como é o caso da linguagem de programação *Python* ou fluxo de trabalho numa tarefa de *Data Science*.

Após maior conforto com essas ferramentas, fez-se um levantamento dos métodos de *machine learning* e de processamento de linguagem natural, com a utilização de uma base de dados para a classificação de texto de uma competição, onde é possível comparar os diferentes modelos.

No terceiro passo fez-se uma análise de como os dados são disponibilizados pelos sistemas de gestão atuais, maneiras de serem importados, análise fiabilística a seguir e métodos de representação dos dados.

Na última etapa criou-se um protótipo do que deverá ser a funcionalidade final de um *software* capaz da análise de ordens de trabalho.

Análogamente, pode ser analisada a metodologia de *Data Science*, que também foi utilizada neste projeto, mais concretamente na terceira e quarta etapas:

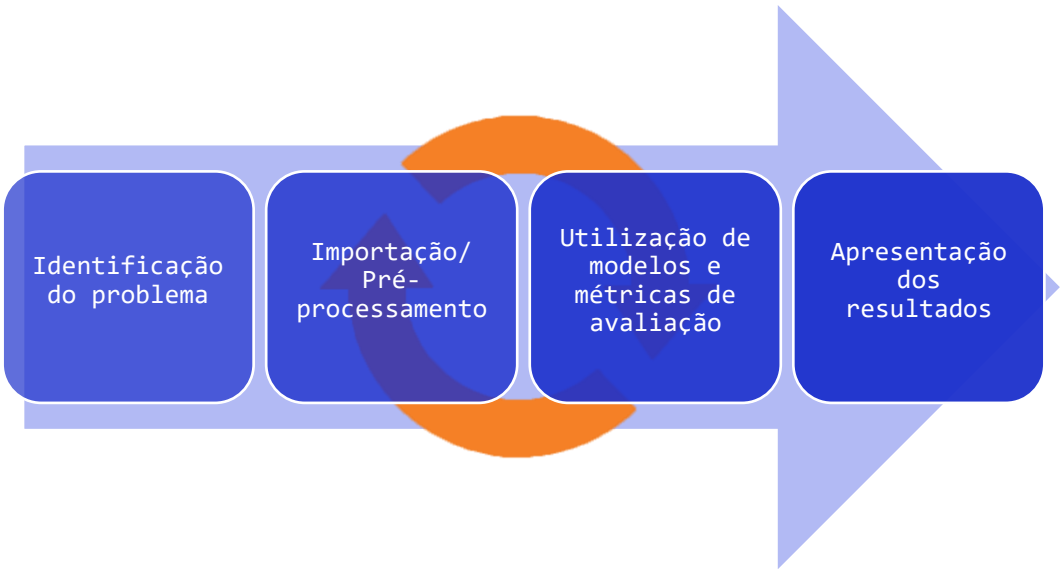


Figura 3 - *Data Science Workflow*

Na identificação do problema serão levantadas as metodologias, objetivos e marcos da tarefa a realizar. Na importação e pré-processamento dos dados, os dados serão tratados de modo a

serem mais facilmente utilizados pelas técnicas e modelos desenvolvidos ou escolhidos para a tarefa. Na utilização dos modelos deverão ser usadas métricas de avaliação como modo de decisão de qual modelo usar para cada situação que surge, e, iterativamente deverá ser feito o pré-processamento para registar as consequências, novamente com métricas de avaliação. Numa fase final, é necessário tratar da apresentação dos resultados obtidos e respetivas conclusões, de um modo mais compacto e pertinente.

1.5 Estrutura da Dissertação

Esta dissertação está dividida em cinco capítulos:

O primeiro capítulo introdutório, que já foi previamente descrito;

O segundo capítulo de enquadramento teórico e análise do estado de arte, onde são descritas todas as práticas, metodologias e algoritmos usados durante o trabalho, desde a fase de processamento de linguagem natural aos resultados fiabilisticos das ordens de trabalho;

O terceiro capítulo de apresentação do problema introduz o conceito de ordem de trabalho; as suas características, ciclo de vida, parâmetros, aplicações e limitações existentes;

O quarto capítulo apresenta a solução encontrada e o protótipo desenvolvido em maior pormenor;

O quinto e último capítulo descreve as conclusões retiradas durante o projeto, e possíveis desenvolvimentos futuros nas áreas envolvidas.

2 Enquadramento Teórico

Neste capítulo será apresentado de forma sucinta a informação que este projeto tem como base. Numa primeira parte irão ser apresentados os processos de extração e tratamento de dados utilizados no processamento de linguagem natural. De seguida será feita uma análise aprofundada dos algoritmos de *machine learning* estudados e utilizados durante o projeto e conclui-se com uma apresentação de algumas técnicas fiabilísticas utilizadas para a análise dos dados tratados.

2.1 Análise e Processamento de Texto

No processo de análise de texto automatizado, por técnicas de *machine learning* ou por outros métodos para o efeito, existem etapas básicas, mas cruciais, que serão expostas de seguida. Após essas etapas, serão necessários mecanismos simplificadores do problema, de modo a que seja possível a análise de sentimento das palavras.

2.1.1 Conceitos Gerais

Tokenização

Processo de normalização de texto, onde são extraídas de um documento as suas palavras constituintes, chamadas de *tokens* ou termos. Existem vários conjuntos de regras ou *standards* para a definição de uma palavra (para além da separação por espaçamentos), visto que pode ser considerada ou desconsiderada a pontuação como hífens na língua portuguesa ou o apóstrofo na língua inglesa, maiúsculas. Um dos métodos mais utilizados é a de *Tokenização* de *Penn Treebank* (Jurafsky e Martin, 2017), onde a pontuação, excetuando o travessão, é considerada um *token* e o travessão é considerado uma ligação entre *tokens*. Este é um método mais conservador, visto que mantém integralmente a escrita das palavras, não alterando, por exemplo, a capitalização das palavras. Métodos mais sofisticados, como o *MaxMatch* (Jurafsky e Martin, 2017), por comparação, encontram os vocábulos com o número máximo de caracteres, existentes no léxico de palavras.

Truncatura

Processo de análise da raiz das palavras, onde são removidos os prefixos e sufixos de uma palavra de modo a aproximar-se ao radical. Este processo está mais explorado na língua inglesa, sendo o *Porter Stemmer* um exemplo simples e eficiente (Jurafsky e Martin, 2017). Este processo é útil para atribuir valores de sentimento semelhantes a palavras que originam do mesmo radical.

Lematização

Processo com mesmo objetivo ao de truncatura, que difere na metodologia. Neste caso, a extração do radical de uma palavra provém de um dicionário de lemas. Este método é mais sofisticado e mais preciso, e é especialmente eficaz para converter tempos verbais no seu infinitvo.

N-gramas

Análise de sequência de N palavras, com analogia à propriedade de *Markov*, segundo Jurafsky e Martin (2017), que assume que um evento futuro apenas depende do presente (neste caso, a palavra futura só depende das N palavras que estão diretamente atrás da mesma). Esta análise permite adicionar maior complexidade, e, consequentemente melhores resultados à análise de sentimento.

2.1.2 Lei de *Zipf*

Segundo Piantadosi (2014), um grande número de fenómenos empíricos seguem a distribuição que é proporcional de acordo com a expressão generalizada de *Mandelbrot* (Equação 2.1),

$$f(r) = \frac{1}{(r+\beta)^{\alpha}} \tag{2.1}$$

Onde:

α e β são constantes e

r é o índice de uma lista decrescente da frequência de um acontecimento

Um desses fenómenos é o da linguagem natural, onde as constantes tomam o valor de $\alpha \approx 1$ e $\beta \approx 2{,}7$. A explicação teórica para esta lei é não trivial, sendo as áreas de conhecimento abrangentes tão divergentes como Estatística e Psicologia. Esta lei também se aplica a outros fenómenos, tais como em programação, *Networks* e música.

Esta lei torna-se relevante quando se identificam as palavras com maior utilização num conjunto de documentos. Visto que estas normalmente são consideradas palavras vazias e não têm impacto na classificação de texto. A título de exemplo, grande parte dos determinantes são considerados como palavras vazias (palavras como ‘o’, ‘a’, ‘meu’, ‘isso’), pois a informação acrescentada pela palavra em si é quase nula, e só quando conjugada com outras palavras se torna relevante.

2.1.3 TF-IDF (Term Frequency-Inverse Document Frequency)

Para a análise e preparação dos dados, é necessário transformar o texto numa quantidade de parâmetros quantificados, e a quantificação deverá seguir um critério que seja ajustado para dar maior relevância a informação importante, de uma forma automatizada.

O critério TF-IDF é um parâmetro para reconhecer termos importantes num texto, relacionando a frequência de um termo num dado documento com a frequência de um termo na totalidade dos documentos. Os termos mais importantes serão os que aparecem com maior frequência num documento e que apareçam em baixa frequência na totalidade dos documentos, e estes irão ter um TF-IDF mais elevado.

Para título ilustrativo, imagine-se um conjunto de documentos que perfazem a população de relatórios de manutenção. Num dado documento existem N menções ao termo “Conjunto Bomba-Motor NºX”, e que este termo apenas aparece num reduzido número de documentos. Isto significa que o termo terá um valor TF-IDF elevado e isto deverá conduzir a uma categorização como “Documento referente ao Conjunto Bomba-Motor NºX”.

O TF (*Term Frequency*) é indicador da frequência de um termo num dado documento, e pode ser expresso de diferentes maneiras, podendo ser booleano, caso seja apenas registada a existência ou não do termo no documento, inteiro, caso seja registada a frequência, ou pode ser

normalizado com o número de termos no documento, com a frequência máxima de um termo registada no documento ou logaritmicamente.

$$\mathbf{TF} = f_{t,d} \tag{2.2}$$

,

$$\mathbf{TF} = log(1 + f_{t,d}) \tag{2.3}$$

ou

$$\mathbf{TF} = 0{,}5 + 0{,}5 \times \frac{f_{t,d}}{max_{t\prime \in d}(f_{t\prime,d})} \tag{2.4}$$

Onde:

$f_{t,d}$ é a frequência do termo t no documento d e

$max_{t\prime \in d}(f_{t\prime,d})$ é a frequência do termo t' com frequência máxima no documento d

A equação 2.2 é uma matriz de frequências, a equação 2.3 reflete uma normalização logarítmica com suavização enquanto que a equação 2.3 é uma normalização dupla com a frequência máxima de um dado documento.

O IDF (*Inverse Document Frequency*) é indicador da raridade de um termo na população dos documentos. Normalmente é calculado pelo logaritmo do quociente do número total de documentos pelo número de documentos que contêm o termo em estudo. Este poderá ser suavizado, para evitar valores infinitos.

$$\mathbf{IDF} = log\left(\frac{N}{n_t}\right) \tag{2.5}$$

ou

$$\mathbf{IDF} = log\left(1 + \frac{N}{n_t}\right) \tag{2.6}$$

Onde:

N é o número total de documentos e

n_t é o número de documentos onde existe o termo t

A equação 2.5 é a matriz da frequência inversa de documentos, enquanto que a equação 2.6 sofre de uma suavização.

O TF-IDF (Equação 2.7) vem então pelo produto de *Hadamard* das duas matrizes,

$$\mathbf{TFIDF} = \mathbf{TF} \circ \mathbf{IDF} \tag{2.7}$$

2.1.4 CBTW (Category-Based Term Weights)

Um dos problemas que se enfrenta na classificação é o desequilíbrio entre classes, isto é, classes que são ou não verificadas em apenas uma minoria da população ou amostra. Para combater uma generalização dos modelos, e consequentemente uma classificação igual para qualquer que seja o caso, foi proposta uma alternativa por Liu et al. (2007), onde termos que aparecessem em classes mais raras terão maior peso para o modelo.

Tabela 1. Elementos de cálculo do CBTW (Liu et al. 2007)

	$\mathbf{c_i}$	$\mathbf{\overline{c_i}}$
$\mathbf{t_k}$	**A**	**B**
$\mathbf{\overline{t_k}}$	**C**	**D**

Onde:

$\mathbf{c_i}$ e $\mathbf{\overline{c_i}}$ impõem a condição do documento pertencer ou não à classe i, respetivamente,

$\mathbf{t_k}$ e $\mathbf{\overline{t_k}}$ impõem a condição do documento ter ou não o termo k, respetivamente e

A, **B**, **C** e **D** o número de documentos com as condições impostas (notar que serão matrizes de dimensão $k \times i$)

É de notar que estes elementos de cálculo necessitam de amostras já classificadas para que sejam calculados.

O CBTW será um grupo de i matrizes de dimensão igual à matriz de frequência de termos normalizada ($d \times k$), e é calculado segundo a equação 2.8, repetindo para cada documento d:

$$\mathbf{cbtw}_{i,d} = \mathbf{ntf}_{d} \circ \log\left(1 + \frac{\mathbf{a_i}}{\mathbf{b_i}} \times \frac{\mathbf{a_i}}{\mathbf{c_i}}\right) \tag{2.8}$$

Onde:

$\mathbf{cbtw}_{i,d}$ é o vetor linha d da matriz i de CBTWs,

$\mathbf{ntf}_{d}$ é o vetor linha d da matriz de frequências normalizadas e

$\mathbf{a}_i$, $\mathbf{b}_i$ e $\mathbf{c}_i$ são vetores coluna i das matrizes A, B e C, respetivamente

Sendo a matriz **NTF** (*Normalized Term Frequency*) calculada segundo a equação 2.9:

$$\mathbf{NTF} = \frac{f_{t,d}}{max_{t' \in d}(f_{t',d})} \tag{2.9}$$

2.2 Modelos de Classificação

Para a análise e tratamento de texto foi necessária a revisão do conhecimento em modelos de *machine learning* para classificação de dados. Os modelos estudados foram os de Naïve Bayes, máquinas de suporte de vetores, árvores de decisão, regressão logística e redes neuronais. Os modelos estudados são todos classificados como supervisionados pois necessitam de *labels*, para que os parâmetros do modelo são ajustados. Métodos não supervisionados dividem as amostras em diferentes grupos sem necessitarem de uma classificação prévia, como é o caso de *Clustering*.

2.2.1 Árvores de Decisão

Árvores de decisão são algoritmos de classificação que, a partir da divisão dos atributos pelos seus valores possíveis, prevê o valor das classes para um dado exemplo. Este modelo tem uma vantagem significativa em relação aos restantes devido à sua facilidade de interpretação e leitura (Bramer, 2016).

As árvores de decisão são criadas a partir de um conjunto de regras de decisão (ou nodos), que deverão visar um atributo. O atributo deverá ser divido em duas ou mais possibilidades (ramos), e esta divisão deverá continuar a ocorrer até que exista um ramo para cada resultado possível. Para a escolha da ordem da divisão de atributos é possível utilizar diferentes metodologias, sendo as mais utilizadas a Entropia, o Índice de Gini e o Qui Quadrado. Estas métricas são utilizadas de modo a reduzir o número de caminhos existentes na árvore de decisão e agilizar o processo de decisão.

Para cada caminho um atributo pode aparecer em apenas um ou nenhum nodo, não podendo ser repetido, o que não poderá ser tão óbvio para árvores de decisão mais complexas. O processo de criação de nodos acaba quando todos os caminhos estão explorados até se atingir um único valor possível de classes ou quando não existem mais atributos para divisão.

Entropia

A entropia é uma medida que transmite a qualidade da informação que cada um dos atributos fornece. Um valor nulo ou perto de zero de entropia demonstra que um atributo é fulcral para o valor da classe, enquanto que um valor alto demonstra que o atributo não tem peso na decisão do valor da classe.

A entropia é dada pela seguinte equação (2.10),

$$H(E) = -\sum_{i=1}^{K} p_i \, log_2 \, p_i \qquad (2.10)$$

Onde:

p_i é a proporção de ocorrências do valor da classe i e

k é o número de classes

Para a criação da árvore de decisão é inicialmente calculado o valor de entropia inicial $E_{inicial}$, e, para cada atributo é calculado a sua entropia. A divisão do nodo ocorre pelo atributo que mais reduz a entropia. A variação $\Delta E = E_{final} - E_{inicial}$ é chamada de ganho de informação, e é expresso em *bits*.

Alternativas à entropia

O índice de *Gini* é uma medida que desempenha um papel igual ao da entropia, mas calculado de um modo mais leve, não utilizando a função de logaritmo (Equação 2.11).

$$Gini(E) = 1 - \sum_{i=1}^{K} {p_i}^2 \tag{2.11}$$

O critério do χ^2(Qui Quadrado; Equação 2.12) é usado para rejeitar ou não uma hipótese nula H_0, de que um atributo não tem influência no valor da classe. Também pode ser calculado para todos os atributos, e após isso escolher o que tem valor mais elevado.

$$\chi^2 = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{\left(\widehat{f_{i,J}} - f_{i,j}\right)}{\widehat{f_{i,J}}} \tag{2.12}$$

O valor esperado $\widehat{f_{i,J}}$ vem seguinte a equação 2.13:

$$\widehat{f_{i,J}} = \frac{\sum_{m=1}^{K} f_{m,j} \times \sum_{n=1}^{L} f_{i,n}}{\sum_{m=1}^{K} \sum_{n=1}^{L} f_{m,n}} \tag{2.13}$$

Onde:

$f_{i,j}$ é a frequência de eventos da classe i e valor de atributo j

2.2.2 Naïve Bayes

O modelo de *Naïve Bayes* baseia-se na definição do teorema de *Bayes* (equação 2.14) para probabilidades condicionadas,

$$P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})} \tag{2.14}$$

Este modelo tem como objetivo calcular a probabilidade de cada classe e escolher a classe que maximiza os atributos da amostra. O objetivo deste modelo pode ser então escrito como um problema de otimização (Equação 2.15) e aplicado ao problema de classificação de texto como:

$$\hat{c} = \text{argmax}\, P(c|d) \tag{2.15}$$

Onde:

$\hat{c}$ é a classe estimada,

c é uma classe e

d é um documento

Esta expressão pode se traduzida literalmente como a classe que otimiza a função de probabilidade de ocorrência da classe c sabendo que ocorreu o documento d. Então o teorema de *Bayes* pode ser reescrito para a classificação de documentos,

$$\hat{c} = \text{argmax}\, P(c|d) = \text{argmax} \frac{P(d|c)P(c)}{P(d)} \tag{2.16}$$

Como esta probabilidade é calculada para cada documento individualmente, a probabilidade de ocorrência do documento $P(d)$ é constante e independente da classe c, e consequentemente irrelevante para a função de maximização. Isto significa que apenas é importante para a classificação o numerador do teorema de *Bayes*.

O documento é dividido nos termos existentes $(f_1, f_2, \dots, f_n)$ e a $P(d|c)$ é desdobrada,

$$P(d|c) = P(f_1, f_2, \dots, f_n|c) \tag{2.17}$$

Esta probabilidade pode ser então desdobrada novamente,

$$P(f_1, f_2, \dots, f_n|c) = P(f_1|c) \cdot P(f_2|c) \cdot \dots \cdot P(f_N|c) \tag{2.18}$$

E cada probabilidade é estimada pela frequência da palavra f_n quando o documento pertence à classe c pelo somatório das frequências de todas as palavras na classe c,

$$\hat{P}(f_k|c) = \frac{n_{f_k,c}}{\sum_{i \in V} n_{f_i,c}} \tag{2.19}$$

Onde:

V é o conjunto de palavras que formam o vocábulo dos documentos

Caso a frequência de uma dada palavra para uma classe seja nula, ou seja, a palavra não existe nos documentos pertencentes à classe, a probabilidade acima seria 0, o que implicaria que era impossível para um documento com essa palavra ser classificado como pertencente a essa classe. Como este efeito é indesejável, é feita uma suavização (Equação 2.20) à estimativa da probabilidade condicionada.

$$\hat{P}(f_k|c) = \frac{n_{f_k,c} + 1}{\sum_{i \in V}(n_{f_i,c} + 1)} \tag{2.20}$$

A metodologia seguida é o cálculo das estimativas de probabilidade, multiplicação das estimativas e comparação para cada classe c.

Este modelo assume, no entanto, duas premissas erradas, que estão na origem do nome do modelo (*Naïve*, significando ingénuo). Na equação 2.18, onde se multiplica todas as probabilidades condicionadas, é assumido que estas probabilidades são independentes, ou seja, a existência de uma palavra é independente das que já estão presentes, e também é assumido que a ordem no qual as palavras aparecem não tem influência no significado da frase (Jurafsky e Martin, 2017). Apesar das assunções, o modelo apresenta bons resultados, especialmente para segmentos de texto mais curtos.

2.2.3 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (SVM – *Support Vector Machine*) tem como objetivo a criação, a partir de um espaço de n dimensões, coincidentes com o número de parâmetros da amostra de tamanho k, um plano capaz de dividir as classes. Após a criação deste plano, a classificação é feita a partir da localização no espaço da amostra a classificar.

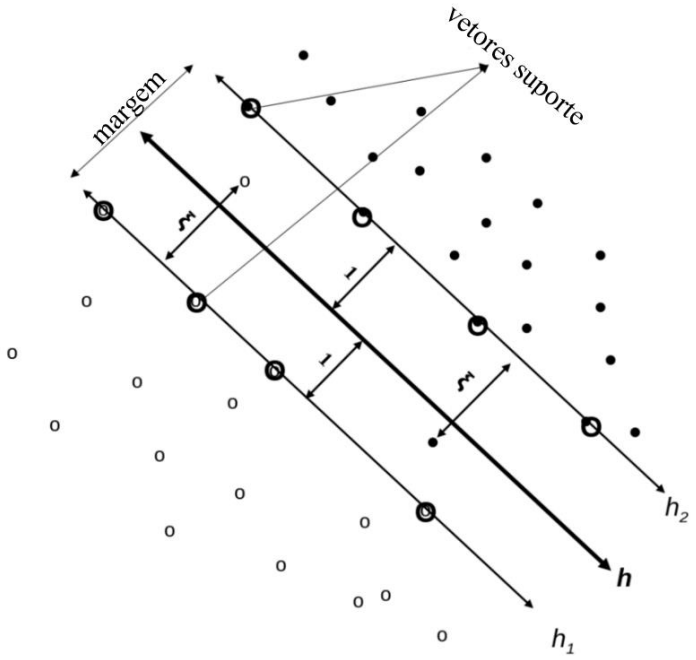


Figura 4 - Exemplo da metodologia das SVM (Tirunagari, 2015)

No caso da figura 1, existem duas dimensões de parâmetros, e as classes são distinguidas por pontos preenchidos ou a branco. O objetivo expresso visualmente é a obtenção de h, plano que divide as duas classes. A partir do plano h é possível deduzir a função de decisão, que vem de $g(\mathbf{x})$ e que toma os valores:

$$sinal\big(g(\mathbf{x})\big) = \begin{cases} +1, g(\mathbf{x}) \geq 0 \\ -1, g(\mathbf{x}) < 0 \end{cases} \quad (2.21)$$

Num caso de variáveis separadas linearmente, a função derivada do plano pode ser definida pela expressão:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \quad (2.22)$$

Onde:

w e b são um vetor normal ao plano e uma constante, respetivamente.

A definição do plano pode tomar infinitos valores, entre h_1 e h_2. A função objetivo é dada pela maximização da margem mínima para todos os valores $\mathbf{x_i}, i = 1, 2, \ldots, k$. A margem pode ser definida:

$$y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b}) \geq 1 \quad (2.23)$$

E o valor da margem será:

$$\gamma = \frac{1}{||\mathbf{w}||} \tag{2.24}$$

Como se trata de um problema de otimização de uma igualdade e desigualdade, é comum utilizar-se o formalismo de *Lagrange* para serem resolvidos (Bishop, 2006):

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}||\mathbf{w}||^2 \sum_{i=1}^{k} a_i\{t_n(\mathbf{w}^T\mathbf{x}_i - b) - 1\} \tag{2.25}$$

Caso existam *outliers* (marcados na figura pela letra ξ), é também necessária uma minimização de função de erro:

$$\Theta(\xi) = \sum_{i=1}^{k} \xi_i \tag{2.26}$$

Para dados que não são linearmente separáveis, o aumento da dimensionalidade através de uma combinação dos atributos (a partir de um núcleo) pode resultar numa solução para a classificação, com redução da função de erro. Os núcleos normalmente são do tipo linear (Equação 2.27), gaussiano (Equação 2.28) ou polinomial (Equação 2.29).

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_k) = \langle \mathbf{x}_i, \mathbf{x}_k \rangle \tag{2.27}$$

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_k) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_k||^2}{\sigma}\right) \tag{2.28}$$

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_k) = (\langle \mathbf{x}_i, \mathbf{x}_k \rangle + c)^k \tag{2.29}$$

2.2.4 Redes Neuronais Artificiais

Uma rede neuronal artificial é um modelo de previsão formado por várias camadas de "neurónios", que enviam sinais entre si, desde a camada inicial de *input*, até à última camada de *output*.

Um neurónio é a unidade básica da rede neuronal artificial. Este recebe um conjunto de sinais provenientes dos neurónios da camada anterior e uma constante (normalmente apelidada de *Bias*), faz o somatório, e com recurso a uma função de ativação, emite o sinal para a próxima camada. As exceções a este modo de funcionamento são a camada de *input* e de *output*. A camada de *input* terá a dimensão do número de parâmetros da amostra e recebe um sinal por parâmetro. A camada de *output* transmite a previsão do modelo. Assim sendo, estas camadas têm um número de neurónios igual ao número de parâmetros e ao número de classes possíveis de classificação.

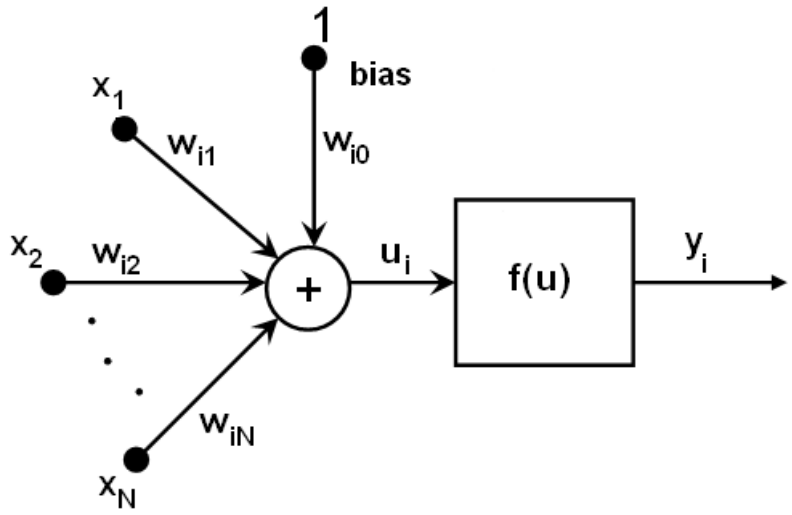


Figura 5 - Formato e função de um neurónio

O funcionamento do neurónio de uma rede neuronal pode ser então descrito pelas equações 2.30 e 2.31:

$$u_i = w_{i0} \sum_{j=1}^{N} (x_j \times w_{ij}) \tag{2.30}$$

Onde:

u_i é o sinal que o neurónio i recebe,

w_{i0} é o bias do neurónio i,

x_j é o sinal que o neurónio j emite e

w_{ij} é o peso do sinal do neurónio j no neurónio i

$$y_i = f(u_i) \tag{2.31}$$

Onde:

u_i é o sinal que o neurónio i recebe calculado pela equação anterior,

f é a função de ativação dos neurónios e

y_i é o sinal que o neurónio i emite

As funções de ativação são funções que normalmente partilham um número deste conjunto de características:

- Passam pela origem, o ponto $(0; 0)$;
- Têm assintotas horizontais;
- Tendem rapidamente para as assintotas horizontais;
- Têm valores de y positivos quando x é positivo e de y negativos quando x é negativo, ou seja, são traçadas no primeiro e terceiro quadrantes.

Exemplos destas funções são a Tangente hiperbólica, Arco de tangente, função identidade, sigmoide e ReLU (função identidade retificada). As que demonstram melhores resultados são as funções de tangente hiperbólica, sigmoide e mais recentemente, a ReLU, descritas nas equações 2.32, 2.33 e 2.34, respetivamente:

$$f(u) = \frac{(e^u - e^{-u})}{(e^u + e^{-u})} \tag{2.32}$$

$$f(u) = \frac{1}{1 + e^{-u}} \tag{2.33}$$

$$f(u) = \begin{cases} u \; se \; u \geq 0 \\ 0 \; se \; u < 0 \end{cases} \tag{2.34}$$

Retropropagação/ método do gradiente

Para a parte de aprendizagem do modelo, é utilizado o método de retropropagação ou método do gradiente, que consiste na minimização do erro de previsão dada uma certa amostra e a classificação correta. Para isso, é necessário derivar parcialmente a função do erro em relação aos neurónios de output, e continuar a derivação em relação às camadas anteriores.

Após a derivação é necessário ajustar os parâmetros do peso e do *bias* de cada neurónio, de modo a encontrar um erro mínimo, ou seja, anular a derivada parcial. Para este método é conveniente que a função de ativação do neurónio seja facilmente derivável, e daí a sua importância. O resultado obtido será um vetor com a direção de declive mais baixo na função do erro, e com sucessivas iterações, é possível encontrar um mínimo local da função de erro de previsão.

2.2.5 NBSVM

Em muitos casos, a combinação de diferentes formas de dois ou mais classificadores permite uma melhor performance do que os classificadores separados. Para a análise de texto, a combinação do modelo de *Naïve Bayes* com máquina de vetores suporte (NBSVM) tem uma performance boa e uniforme em diferentes tarefas e bases de dados (Wang e Manning, 2012).

Tal como no modelo SVM, NBSVM será um classificador linear, seguindo a equação (2.21) e (2.22) para a previsão. No entanto os parâmetros introduzidos para o modelo serão substituídos pelos parâmetros de *Naïve Bayes*.

Para todos os parâmetros é criado um vetor **r** de rácio entre frequências das duas classes:

$$\mathbf{r} = \log\left(\frac{\mathbf{p}/\|\mathbf{p}\|_1}{\mathbf{q}/\|\mathbf{q}\|_1}\right) \tag{2.35}$$

Em que o vetor **p** e o vetor **q** são definidos da seguinte maneira:

$$\mathbf{p} = \sum_{i:y^{(i)}=1} \mathbf{f}^{(i)} \tag{2.36}$$

$$\mathbf{q} = \sum_{i:y^{(i)}=-1} \mathbf{f}^{(i)} \tag{2.37}$$

Onde:

$\mathbf{f}^{(i)}$ é o número de ocorrências de um dado parâmetro na amostra i e

$y^{(i)}$ é a classificação da amostra i

Com a definição destas novas variáveis, a minimização que ocorre no modelo SVM pode ser reescrita:

$$\mathbf{w}^T\mathbf{w} + C\sum_i \max\left(0, 1 - y^{(i)}\left(\mathbf{w}^T\mathbf{f}^{(i)} + b\right)\right)^2 \tag{2.38}$$

A alteração deste classificador ocorre na variável de entrada **x** que é definida de modo diferente, calculada da seguinte maneira:

$$\mathbf{x}^{(k)} = \tilde{\mathbf{f}}^{(k)} = \hat{\mathbf{r}} \circ \hat{\mathbf{f}}^{(k)} \tag{2.39}$$

Onde:

$\hat{\mathbf{r}}$ e $\hat{\mathbf{f}}^{(k)}$ são os valores estimados de rácio e de frequência para a população k, deduzidos da amostra de treino

Para adicionar mais robustez ao modelo, é feita uma interpolação no vetor **w**, $\mathbf{w}'$:

$$\mathbf{w}' = (1-\beta)\bar{\omega} + \beta\mathbf{w} \tag{2.40}$$

Onde:

$\bar{\omega}$ é a magnitude média de **w** e

β é o parâmetro de interpolação

Segundo Wang e Manning (2012), valores ótimos para o parâmetro β situam-se no intervalo $\left[\frac{1}{4}, \frac{1}{2}\right]$.

2.3 Métricas de Desempenho de um Modelo

Para a escolha mais correta do modelo que se pretende usar, é necessária uma análise à performance do mesmo em relação aos dados fornecidos. Para isso, foram criadas métricas de avaliação dos modelos, que vão ser apresentadas de seguida.

2.3.1 Indicadores Gerais de um Modelo de Classificação

Numa classificação binária, onde exista um desequilíbrio dos valores das classes, existe uma tendência para o modelo se adaptar apenas aos valores dominantes, classificando corretamente a maioria das amostras que pertencem ao valor dominante, mas errando as amostras que são do valor minoritário. Isto pode resultar em modelos que, apesar da elevada exatidão, não são realmente uteis.

Segundo Tirunagari (2015), o recurso a uma matriz de confusão é mais informativo quanto ao funcionamento do modelo. Assumindo uma classificação positiva/negativa, a matriz de confusão vem do seguinte modo:

Tabela 2 - Matriz de confusão de um problema de classificação binária

	p' **(valor previsto)**	**n'** **(valor previsto)**
p **(valor real)**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
n **(valor real)**	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Com esta matriz é possível calcular outras medidas mais sensíveis às classes minoritárias. Segundo esta notação, a exatidão do modelo é calculada:

$$ACC = \frac{TP + TN}{p + n} \tag{2.41}$$

Uma das medidas mais usadas para conhecer a performance do classificador é o *F measure*, que relaciona a precisão e a sensibilidade, que tem como objetivo dar importância a classificação de classes minoritárias.

$$F = 2 \times \frac{precisão \times sensibilidade}{precisão + sensibilidade} \tag{2.42}$$

Onde:

$precisão$ é definida por $Pr = \frac{TP}{TP+FP}$

$sensibilidade$ é definida por $Re = \frac{TP}{TP+FN}$

2.3.2 Curva Característica de Operação do Recetor

A curva característica de operação do recetor, normalmente chamada de curva ROC, é um gráfico de dispersão da sensibilidade com o rácio de falsos positivos $FPR = \frac{FP}{FP+TN}$, sendo a sensibilidade o eixo vertical e o rácio de falsos positivos o eixo horizontal. O resultado do modelo numa amostra de teste é representado no gráfico por um ponto, e a variação do limiar cria a curva ROC. Os modelos têm como output um sinal ou conjunto de sinais que variam entre 0 e 1 (resultado de confiança). O limiar é o valor que divide os sinais considerados 0 e 1, na classificação binária. Os vértices (0,1) e (1,0) correspondem a um classificador ótimo e um classificador que apresenta os resultados sempre errados, respetivamente; os vértices (0,0) e (1,1) correspondem a classificadores que apresenta os resultados sempre negativos e positivos, respetivamente (Runkler, 2016).

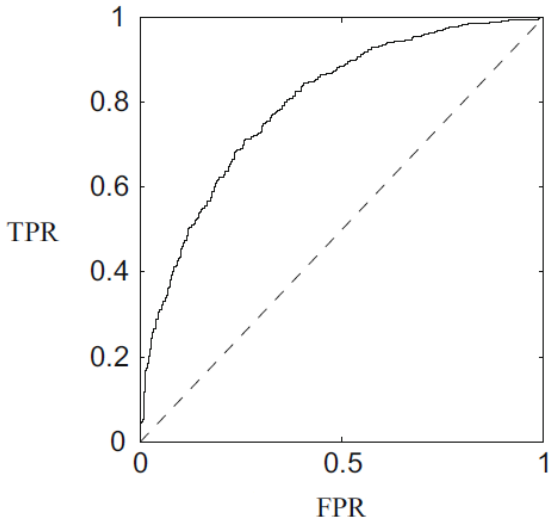


Figura 6 – Representação de uma curva ROC genérica (Runkler, 2016)

A traço interrompido está representada a reta de "geração aleatória" de resultados com uma dada proporção para valores positivos e negativos. Um ponto que esteja abaixo desta reta terá um resultado melhor caso as previsões sejam alternadas para o valor inverso, isto é, os valores positivos passarem a negativos e vice-versa.

Para a avaliação do desempenho de um modelo, é calculada a área debaixo da curva ROC, ROC AUC. O valor máximo atingido para esta métrica é de 1, no caso de o modelo conseguir dividir perfeitamente as classes, e atingir o vértice (0,1), e nunca deverá ir abaixo da área de geração aleatória, 0,5.

Após traçar a curva, a partir da variação do limiar, esta pode ser suavizada, e para obter a sua área, utilizar o método de *Simpson*, ou aproximar a função e integrar. Esta métrica relaciona-se com o índice de Gini, que pode ser definido como o dobro da área dada pela curva e a reta de geração aleatória (Hand e Till, 2001).

$$Gini = 2 \times AUC - 1 \tag{2.43}$$

2.3.3 Validação Cruzada

O método de validação cruzada é usado para estimar com as amostras de treino o desempenho do modelo, tendo em conta as métricas referidas anteriormente. O método consiste na divisão em k partes de igual tamanho, aleatoriamente distribuídas.

Existe um compromisso entre um valor de k baixo, que provoca uma amostra mais baixa para treino do modelo e consequentemente um modelo menos semelhante ao final, e um valor de k alto provoca uma amostra de teste mais baixa e consequentemente métricas menos precisas. É consenso que um valor de $k = 10$ é um bom compromisso entre estes fatores, usando 90% das amostras para treino e 10% para teste. (Tirunagari, 2015)

São alternativas a este método métodos de reamostragem, como o *Bootstraping* ou *Jackkniffing*.

2.4 Análise de Fiabilidade

O objetivo deste trabalho será criar um algoritmo capaz de detetar a partir do texto diferentes modos de falha, e de acordo com os dados temporais, caracterizar as respetivas funções de sobrevivência.

Para isso, inicialmente será revisto o teste de tendência de *Laplace*, que caracterizará o estado do processo, seguido de métodos de parametrização de distribuições a partir dos tempos de falha.

2.4.1 Teste de *Laplace*

O teste de hipótese de *Laplace* é um teste de hipóteses formulado de modo a determinar se um determinado acontecimento pode ou não ser considerado um processo de Poisson homogéneo (Assis, 2004). Para isso são criadas a hipótese nula H_0 de que os tempos entre os acontecimentos são independentes entre si e igualmente distribuídos e a hipótese alternativa H_1 de que os tempos não são independentes ou não seguem a mesma distribuição.

A estatística de teste Z_T é calculada a partir da expressão, quando a limitação é dada pelo tempo:

$$Z_T = \sqrt{12 \times N}\left(\frac{\sum_{i=1}^{N} t_i}{N \times T_0} - 0{,}5\right) \tag{2.44}$$

Onde:

t_i é o tempo t do acontecimento da ordem i,

Né o número de acontecimentos e

T_0 é o tempo acumulado no final do teste

Quando limitada pelo número de falhas, a expressão é alterada:

$$Z_T = \sqrt{12 \times (N-1)}\left(\frac{\sum_{i=1}^{N-1} t_i}{(N-1) \times T_N} - 0{,}5\right) \tag{2.45}$$

Onde:

T_N é o tempo acumulado da última avaria

Pressupondo que os acontecimentos seguem processo de *Poisson*, o valor da estatística de teste tende a uma distribuição normal padronizada $Z \approx N(0{,}1)$, quando o número de acontecimentos é igual ou superior a 4 (Assis, 2004). Assim, definindo um nível de significância α, obtêm-se os valores críticos para a qual a hipótese nula será rejeitada, $Z_{\alpha/2}$ e $-Z_{\alpha/2}$.

Valores de Z_T superiores ao valor crítico positivo indicarão uma tendência crescente, o que significa em termos fiabilísticos, um decréscimo da taxa de avarias; em sentido contrário, valores inferiores ao valor crítico negativo indicarão uma tendência negativa, ou seja, um aumento a taxa de avarias. Para valores dentro do intervalo dos valores críticos, a hipótese nula não poderá ser rejeitada, e o processo deverá ser considerado de Poisson homogéneo.

2.4.2 Método dos Mínimos Quadrados

O método dos mínimos quadrados é um método de otimização que procura encontrar os parâmetros de uma função genérica para a ajustar a um conjunto de dados, minimizando a soma dos erros absolutos, S.

O erro é definido pela diferença do valor das ordenadas obtidas pela função $f(x_i, \theta_1, \theta_2, \ldots, \theta_K)$ com o valor real y_i:

$$r_i = y_i - f(x_i, \theta_1, \theta_2, \ldots, \theta_K) \tag{2.46}$$

E o erro absoluto é dado então pela equação:

$$S = \sum_{i=1}^{n} r_i \tag{2.47}$$

No caso de ajuste dos dados a uma distribuição, a metodologia sugerida é a de calcular a probabilidade cumulativa empírica com o conjunto de dados, com a soma de frequências dos valores anteriores:

$$y_i = \frac{\sum_{j=1}^{i} f_j}{\sum_{j=1}^{N} f_j} \tag{2.48}$$

Onde:

f_j é a frequência do valor j e

N é o conjunto de todos os dados

Após a criação da função empírica o erro será calculado diretamente pelas equações 2.46 e 2.47.

Para o cálculo dos parâmetros, deverá ser feita uma derivação parcial da função do erro em relação ao parâmetro a otimizar. Para se obter o mínimo de erro, as funções derivadas deverão ser anuladas, por métodos analíticos ou por métodos numéricos quando necessário.

Um dos algoritmos utilizados para soluções numéricas é o método do gradiente, que localiza o mínimo local a partir do declive máximo de uma função de K parâmetros.

2.4.3 Método da Máxima Verosimilhança

O método de máxima verosimilhança (MLE) é um método utilizado para a estimação de K parâmetros θ_K, de modo a ajustar um dado modelo estatístico f, para uma amostra de tamanho N, X_N (Lee e Wang, 2003). A metodologia inicia-se ao definir uma função de verosimilhança, que é literalmente a função de densidade de probabilidade do modelo para a amostra X_N condicionada pelos parâmetros θ_K.

$$L(\theta) = \prod_{n=1}^{N} f(X_n, \theta_1, \theta_2, \ldots, \theta_K) \tag{2.49}$$

O objetivo será encontrar o conjunto de parâmetros $\widehat{\theta_K}$ que melhor aproximam o modelo estatístico à distribuição, ou seja, que maximizem o valor da função de verosimilhança. Para simplificação, utiliza-se o logaritmo desta função, $l(\theta)$.

$$l(\theta) = \log\big(L(\theta)\big) = \log\left(\sum_{n=1}^{N} f(X_n, \theta_1, \theta_2, \ldots, \theta_K)\right) \tag{2.50}$$

O máximo é encontrado anulando a derivada parcial da função de verosimilhança com cada um dos parâmetros.

$$\frac{\partial l}{\partial \theta_k} = 0 (k = 1, 2, \dots, K) \tag{2.51}$$

O resultado será um conjunto de K equações com K incógnitas que deverão ser resolvidas em função da amostra X_N. A solução das equações poderá ser obtida por métodos analíticos ou numéricos, como o método de *Newton—Raphson* (Lee e Wang, 2003), dependendo da complexidade do modelo estatístico e da quantidade de parâmetros do modelo.

No caso da existência de amostras censuradas, será necessária a alteração da função de verosimilhança. As amostras censuradas não apresentam o valor da característica medida, mas sim um intervalo de valores possíveis, e, para a medição da probabilidade, não será usada a função de densidade de probabilidade f, mas sim a função de distribuição cumulativa F. Assim, para cada amostra censurada, será medida a probabilidade de a característica pertencer ao intervalo, dado o modelo estatístico e os respetivos parâmetros.

A função genérica de verosimilhança contempla um conjunto de amostras não censuradas X_N e uma amostra censurada à esquerda, direita ou por intervalo X_o.

$$L(\theta) = \prod_{n=1}^{N} f(X_n, \theta_1, \theta_2, \dots, \theta_K) \times F(X_o, \theta_1, \theta_2, \dots, \theta_K) \tag{2.52}$$

Como X_o pode tomar três formas, a função F é definida também de forma diferente para cada tipo de amostra.

Tabela 3 - Cálculo da probabilidade de amostras censuradas

Censurada à esquerda	$\boldsymbol{X_o =]-\infty, a]}$	$\boldsymbol{F(X_o) = F(a)}$
Censurada à direita	$X_o = [b, +\infty[$	$F(X_o) = 1 - F(b)$
Censurada em intervalo	$X_o = [c, d]$	$F(X_o) = F(d) - F(c)$

O procedimento após a criação da função de verosimilhança é igual ao caso de amostras não censuradas.

2.4.4 Convolução de Distribuições

Uma convolução é um método utilizado para definir uma nova distribuição Z, resultante da soma de duas distribuições X e Y (Bolch et al., 2006). Esta soma pode ser expressa pela equação 2.53:

$$\mathrm{Z} = \mathrm{X} + \mathrm{Y} \tag{2.53}$$

Onde:

Z, X *e* Y são 3 variáveis aleatórias (e seguem cada uma a sua distribuição)

O cálculo da função de densidade de probabilidade da variável Z, $f_z(z)$ é calculado a partir das funções de densidade de probabilidade das variáveis originais:

$$f_z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx, \qquad -\infty < z < +\infty \tag{2.54}$$

O domínio do integral pode ser alterado, caso as funções não estejam definidas nesses valores (por exemplo, o integral da soma de duas funções exponenciais negativas deverá ser feito em $[0, +\infty[$).

Este método usufrui a propriedade comutativa, o que significa que a integração pode ser feita em ordem a y.

Os exemplos mais comuns deste método é a soma de duas distribuições normais, que resulta numa distribuição normal com parâmetros iguais à soma das médias e das variâncias das distribuições anteriores; a soma de duas distribuições uniformes resulta numa distribuição triangular; a soma de duas distribuições exponenciais negativas resulta numa distribuição hypoexponencial no resultado generalizado, e caso o parâmetro λ seja igual para as duas funções originais, uma distribuição de *Erlang*.

3 Caracterização e Análise de Ordens de Trabalho

Neste capítulo será definida uma ordem de trabalho. Será também apresentado o percurso feito durante a criação e análise de uma ordem de trabalho, as metodologias seguidas, os recursos e plataformas usados, e, posteriormente, será feito um levantamento de potenciais problemas e possíveis melhorias nos processos, quando analisando ordens de trabalho.

3.1 Conceito e Utilização

Uma ordem de trabalho é o elemento básico de registos de operações de gestão da manutenção. Numa ordem de trabalho é realizada uma aglomeração de toda a informação referente a um trabalho de manutenção, tanto a nível de custos previstos e reais (de mão de obra, componentes e materiais, serviços) como a equipamento e componentes envolvidos, tarefas realizadas, instruções necessárias, diagnósticos ou alarmísticas e tipo de ordem de trabalho.

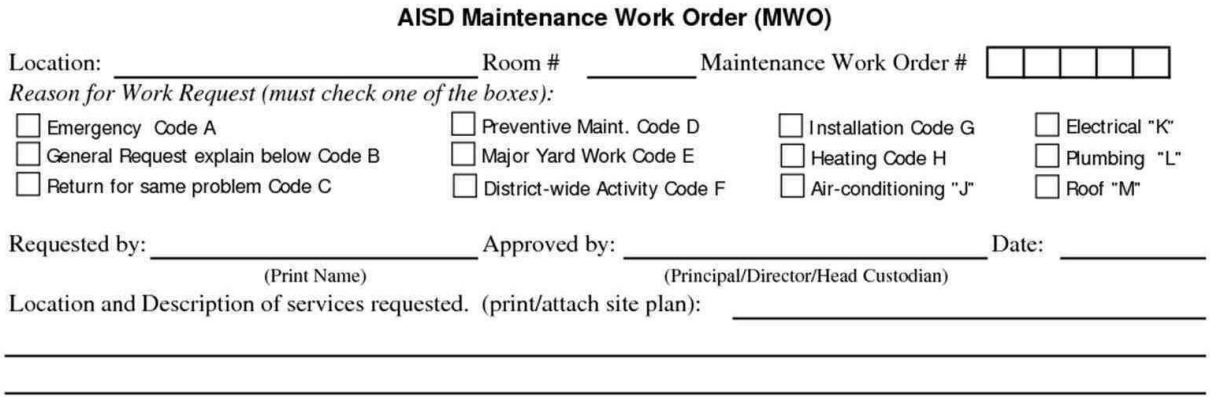
A utilização das ordens de trabalho varia de situação para situação; depende da informação contida na ordem de trabalho e também da política de manutenção da empresa. Em certos casos, pode não existir informação sobre a ordem de trabalho ou esta pode conter pouca informação, enquanto que noutros, a complexidade do documento é tal que a burocracia envolvida diminui a eficácia da utilização de ordens de trabalho e dificultando o trabalho dos técnicos de manutenção, sendo necessário um compromisso na complexidade e na informação contida.

A análise posterior da ordem de trabalho varia também de situação para situação, e mesmo dentro da mesma empresa podem existir diferentes modelos com diferentes níveis de informação, para equipamentos mais ou menos críticos.

De um ponto de vista de manutenção preditiva, onde a monitorização é uma peça fundamental, a informação contida numa ordem de trabalho de inspeção é crucial; na análise de fiabilidade de um dado equipamento para os seus modos de falha, é necessário analisar uma ordem de trabalho para identificar os diferentes modos e registar os seus respetivos tempos de falha; na manutenção preventiva, é possível retirar indicadores de desempenho, como por exemplo o rácio entre operações corretivas e preventivas; é possível registar a duração de uma operação de manutenção, para o cálculo de *Mean Time to Repair* (MTTR);além disso para uma análise de causa-raiz de um acontecimento, as ordens de trabalho servem como um histórico do equipamento ou equipamentos visados.

3.2 Pedido de Trabalho

Inicialmente, quando um aviso ou avaria são detetados, é enviado um pedido de trabalho à equipa de manutenção ou outra entidade responsável pela comunicação deste tipo de situações. Este pedido consiste num formulário onde são explicitados uma série de parâmetros variáveis, que normalmente contem dados como a data e hora do acontecimento, uma breve descrição da anomalia (conteúdo), o equipamento ou componente visado (coordenadas do objeto), o departamento proveniente, a identificação do responsável (coordenadas do pedido), a urgência do pedido e a funcionalidade do equipamento após o acontecimento (parâmetros para gestão).



AISD Maintenance Work Order (MWO)

Location: _____ Room # _____ Maintenance Work Order # ☐☐☐☐☐

Reason for Work Request (must check one of the boxes):

☐ Emergency Code A	☐ Preventive Maint. Code D	☐ Installation Code G	☐ Electrical "K"
☐ General Request explain below Code B	☐ Major Yard Work Code E	☐ Heating Code H	☐ Plumbing "L"
☐ Return for same problem Code C	☐ District-wide Activity Code F	☐ Air-conditioning "J"	☐ Roof "M"

Requested by: _____ Approved by: _____ Date: _____

(Print Name) (Principal/Director/Head Custodian)

Location and Description of services requested. (print/attach site plan): _____

Figura 7 - Exemplo de um pedido

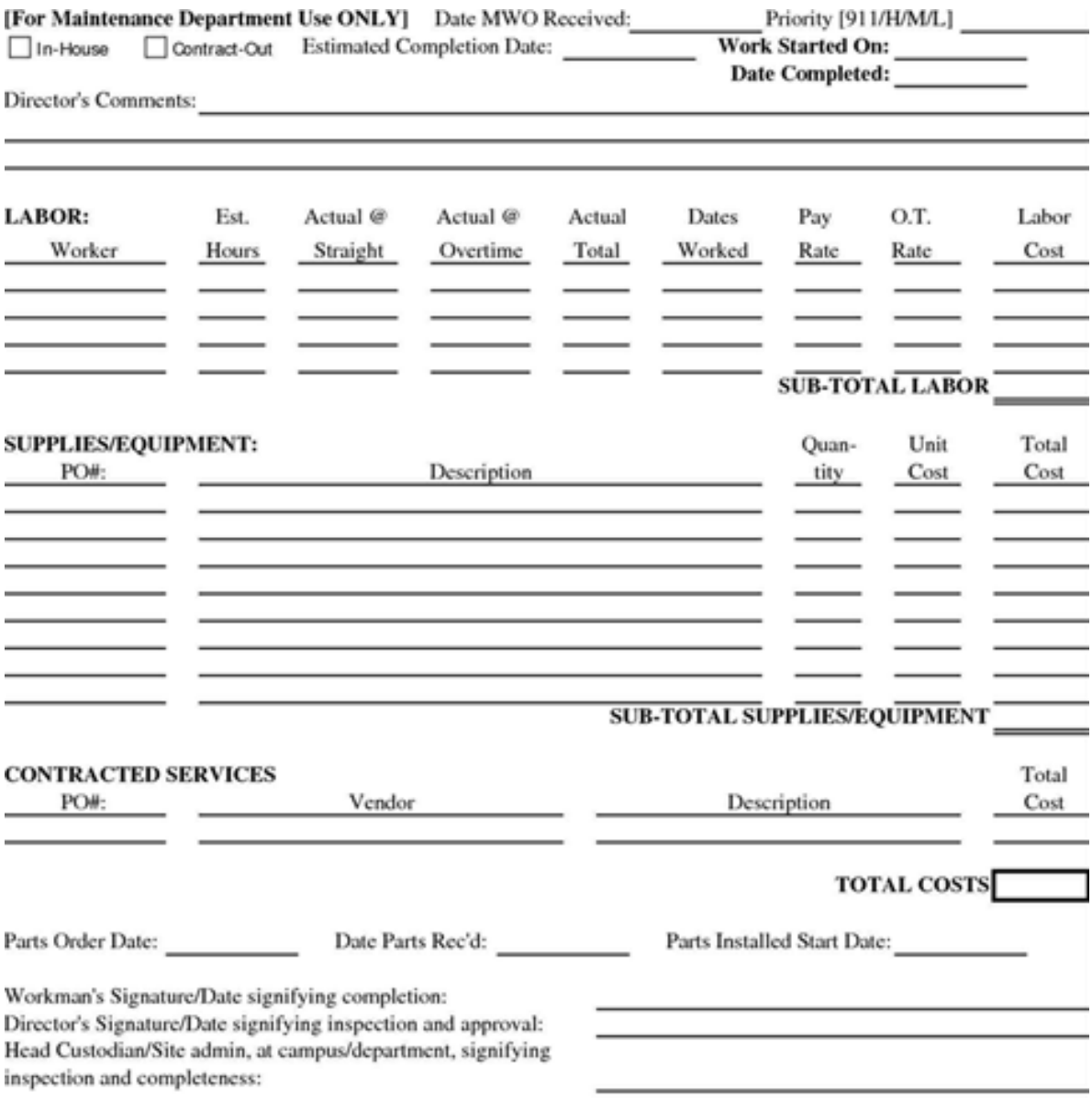
Neste pedido é possível observar alguns dos parâmetros referidos anteriormente, no entanto com variações que se moldam ao objetivo final do documento. Esta ordem de trabalho está adaptada a serviços de manutenção de edifícios, onde os tipos de problemas estão divididos em problemas elétricos, de climatização, da instalação, entre outros.

3.3 Ordem de Trabalho

Após análise de um pedido de trabalho, este será ou não tratado pela equipa de manutenção. Os pedidos tratados emitirão uma ordem de trabalho, que é um outro documento ao qual estará associado novamente uma descrição do sucedido, agora feita pelo técnico de manutenção, componente ou componentes que foram substituídos (caso necessário), data e hora do começo e fim da operação, trabalhador ou trabalhadores envolvidos, custos previstos e reais da operação, entre outros.

Uma ordem de trabalho pode ter várias origens, e não apenas trabalhos corretivos. As origens possíveis das ordens de trabalho, são sistemáticas, de calibração, de rotinas de inspeção e lubrificação, inspeções e condicionada, análise de óleos, preventiva condicional, corretiva ou de melhoria (Cabral, 2006).

A estas operações de manutenção pode estar associada também uma lista de etapas que o técnico de manutenção deve cumprir.



[For Maintenance Department Use ONLY] Date MWO Received: _____ Priority [911/H/M/L] _____

☐ In-House ☐ Contract-Out Estimated Completion Date: _____ **Work Started On:** _____

Date Completed: _____

Director's Comments: _____

LABOR: Worker	Est. Hours	Actual @ Straight	Actual @ Overtime	Actual Total	Dates Worked	Pay Rate	O.T. Rate	Labor Cost
							SUB-TOTAL LABOR	

SUPPLIES/EQUIPMENT: PO#:	Description	Quan-tity	Unit Cost	Total Cost
			SUB-TOTAL SUPPLIES/EQUIPMENT	

CONTRACTED SERVICES PO#:	Vendor	Description	Total Cost

TOTAL COSTS ☐

Parts Order Date: _____ Date Parts Rec'd: _____ Parts Installed Start Date: _____

Workman's Signature/Date signifying completion: _____

Director's Signature/Date signifying inspection and approval: _____

Head Custodian/Site admin, at campus/department, signifying inspection and completeness: _____

Figura 8 - Exemplo de uma ordem de trabalho

Esta ordem de trabalho, por exemplo, está ajustada para o cálculo de custos de uma operação de manutenção. Apesar disso, apresenta informação relativa a outros aspetos anteriormente referidos, tais como a compra de componentes ou equipamentos, data de inicio e fim de reparação e necessidade de subcontratação. Além disso, não aparenta contemplar operações de substituição preventiva, pois não existe campos de preenchimento para custos previstos de operação, ou seja, todas as operações de manutenção deverão ser corretivas.

3.4 Prioridade de uma Ordem de Trabalho

No caso de existência de vários pedidos, será necessária uma escolha de quais as ordens de trabalho que apresentam maior importância e consequentemente serão abordadas em primeiro lugar.

Segundo Cabral (2006), dependendo da ocorrência, o grau de urgência de uma ordem de trabalho deve ser dividido em quatro níveis, U:

1. Emergência – trabalhos corretivos, com risco de segurança e de propagação da falha para grandes proporções.
2. Urgência – trabalhos corretivos ou preventivos necessários para a não existência na baixa de produção.
3. Normal – trabalhos preventivos planeados e agendados; inspeções.
4. Quando conveniente – trabalhos sem impacto no funcionamento das máquinas; cosméticos.

Um aspeto importante é também o equipamento ou componente visado por esta ordem. Estes podem também ser divididos em quatro categorias, determinando a sua criticidade, C:

1. Muito críticos
2. Críticos
3. Normal
4. Baixo

É também necessário ter em conta o impacto do equipamento no funcionamento normal da produção, com o nível hierárquico do solicitante, H:

1. Gestão de topo
2. Produção
3. Gestão Intermédia
4. Outros

A fórmula de prioridade P vem da multiplicação das 3 parcelas:

$$P = U \times C \times H \tag{3.1}$$

Este valor para cada ordem de trabalho encontrar-se-á entre 1 e 64, sendo os valores mais próximos de 1 os de maior prioridade.

3.5 Tratamento dos dados das Ordens de Trabalho

Após o preenchimento dos documentos pelos intervenientes da operação de manutenção, numa etapa de análise e, este é transcrito para um *software* CMMS (*Computerized Maintenance Management System*) como o IBM Maximo, um *software* de gestão como o SAP, ou para uma folha de cálculo ou base de dados, onde a informação será armazenada e eventualmente tratada e analisada, em forma de relatório de trabalhos.

Softwares mais recentes permitem o preenchimento automático de alguns campos de um relatório de trabalhos, como é o caso dos temporais (o tempo de manutenção, tempo de inoperacionalidade, tempo de espera e período de intervenção).

A passagem para uma plataforma de análise possibilita ao utilizador uma visualização histórica de cada um dos elementos intervenientes na operação de manutenção; podem ser criados

relatórios específicos a equipamentos ou componentes, centros de custo ou unidades de produção, técnicos de manutenção, resumos das operações, semanas ou meses; e nesses casos podem ser mostradas as variáveis mais relevantes, o tempo de trabalho, o custo associado à operação ou o tipo de operação.

3.6 Limitações e Consequências da Metodologia Atual

A metodologia atual é um modo rico e completo de registar as ocorrências de manutenção, abordando os pontos cruciais para uma análise fiabilística dos equipamentos, bem como a eficácia dos planos de manutenção postos em prática. No entanto, é permeável a falhas, visto que está dependente do funcionário que executa o pedido e o trabalho de manutenção todo o processo de descrição da ocorrência, diagnóstico e inspeção. Isto significa que partes importantes para a identificação de um modo de falha podem não ser explicitados, e inclusivamente o mesmo modo de falha pode ser descrito de uma maneira diferente por diferentes trabalhadores, alterando o diagnóstico inicial e análise consequente.

Em termos de complexidade, este processo pode ter uma abordagem de um *Template* de ordem de trabalho independentemente da ocorrência a ser resolvida e menos complexa, o que implica mais campos escritos à mão, maior possibilidade de impercetibilidade e erros de transcrição ou ortográficos e menor controlo do processo; pode também levar uma abordagem de grande complexidade, onde para cada departamento ou equipamento exista um tipo de ordem de trabalho diferente, que aumente consequentemente a carga no sistema de gestão e de emissão de ordens de trabalho e também sendo mais rígido a alterações.

Os sistemas de informação utilizados estão otimizados para registar a informação relativa a tempos, nomes de equipamentos, componentes, trabalhadores e custos, mas a parte descritiva do processo não apresenta qualquer valor para estes; uma quantidade significativa da informação contida na ordem de trabalho é apenas utilizada para, por exemplo, análise de modos de falha ou *root cause analysis*, trabalhos executados manualmente pelos analistas da empresa.

3.7 Dados Utilizados

Neste projeto foram utilizados três ficheiros dados, que se encontram em anexo (Anexo A). Estes foram escolhidos pois apresentam formados diferentes, e consequentemente o tratamento dos dados e a extensão da análise possível é diferente.

Os primeiros dois *datasets* foram providos pela EQS; o primeiro é um relatório de trabalho referente a um equipamento (um triturador), proveniente de um *software* de gestão, em forma de ficheiro de texto (*.txt).

Este documento apresenta os seguintes campos relevantes para o projeto:

- Classificação de ordem de trabalho (Cl.) – Campo onde é explicito o tipo de manutenção da OT; vem no formato "COR1" ou "PRV1" para uma operação corretiva ou preventiva respetivamente.
- Inicio e fim da operação (Inic.prog.; Fim.prog.) – Campo onde se apresentam em forma de data o inicio e o fim do trabalho de manutenção (DD.MM.AAAA).
- ID do trabalho (Ordem).
- ID Objeto de gestão (Identificação técnica) – Identificação técnica do equipamento e componente em forma de árvore hierárquica.
- ID da equipa de manutenção (Equipa).
- Nome do objeto de gestão (Denominação da identificação técnica).
- Descrição (Texto breve) – Descrição proveniente da OT, criada pelo técnico de manutenção.
- Custos de operação planeado e real (TotGenPlan; TotGenReal).

Este conjunto de dados é constituído por 146 registos.

O segundo *dataset* fornecido é referente a um conjunto de quinze equipamentos idênticos, proveniente do SAP. Este foi exportado e encontra-se em formato excel (*.xlsx).

Descrição Longa Log	Estado	Tipo
Foi verificada a UPS e no local está sem alarmes. Solicita-se verificação / reparação. Date: 11/15/10 Time: 15:44:08 GMT	FECHADA	MC
Depois de verificações a varios sistemas de segurança concluíu-se que o automato estava bloqueado. Colocou-se o me	FECHADA	MC
NULL	FECHADA	MC
Foram verificadas as duas pontes 146.1 e146.2 estando as duas em parque 07:32. Foi informada a coordenação. OT fechad	REALIZ	MC
foi dada alguma informação de como proceder ficando bem encostada	FECHADA	MC
Conforme falado junto envio fotografia. Cumprimentos, Tomaz Magalhães Crespo.	FECHADA	MC
Esta HOT LINE é testada 2 vezes por dia pelo pessoal do SLCI. Para os SLCI trata-se da linha 116 e hoje quando foi testada	FECHADA	MC
Foi desligada a alimentação e retirados os 3 conjuntos de fusíveis que dão os equipamentos. Foi verificada a não existênd	FECHADA	MC
From: Paulo A. Fradique Sent: domingo, 15 de Agosto de 2010 14:40 To: Ademar C. Lança; ALS CTC-Centro Técnico Cod	FECHADA	MC
Solicitou-se verificação / correcção ao TEA de serviço. (o colega Pedro Pinto retirou o Rádio para reparação).	FECHADA	MC
NULL	FECHADA	MC
Foi ligada a UTA 3	FECHADA	MC
Quando la chegamos verificamos que estava tudo operacional.	FECHADA	MC
Solicita-se correcção/verificação da porta 5.14.204	FECHADA	MC
Foram feitos testes ao fol metálico ao que a operadora adicionou os estalos que tinha ouvido aquele que estava a ouvir	REALIZ	MC
NULL	EMEXEC	MPR

Figura 9 -Amostra dos parâmetros e registos do conjunto de equipamentos

Estes dados apresentam uma estrutura semelhante ao conjunto anterior, não apresentando informação relativa ao modo de falha. Em contrapartida, existem três campos de descrição, e esta encontra-se mais completa, e também existe um campo de prioridade da OT.

Este documento tem uma dimensão maior, apresentado mais de 8000 registos.

O último conjunto de dados foi fornecido pela *Carnegie Mellon University*, inicialmente para utilização numa competição de *machine learning*, em formato de *comma-separated values* (*.csv).

Estes dados apresentam uma estrutura dividida em cinco folhas de cálculo diferentes:

- A primeira folha (machines.csv) contém dados relativos aos equipamentos (ID, modelo e tempo de aquisição)
- A segunda folha (errors.csv) contém informação relativa a alarmística dos equipamentos (data, ID do equipamento, ID do alarme)
- A terceira folha (failures.csv) descreve as falhas dos equipamentos (data, ID do equipamento, componente que falhou)
- A quarta folha (maint.csv) descreve as operações de manutenção realizadas (data, ID do equipamento, componente substituído, tipo de manutenção)
- A última folha (telemetry.csv) é referente a medições de sensores de cada equipamento (data, ID do equipamento, Voltagem, Velocidade de rotação, Pressão, Vibração)

Este representa o conjunto de dados de maior dimensão, tendo a folha de medições uma dimensão superior a 8.000.000 de registos.

Estes registos são referentes a um ano de medições, de mil equipamentos de quatro modelos diferentes.

4 Apresentação da Solução

Neste capítulo será feita a apresentação dos protótipos desenvolvidos durante o projeto, nas fases de importação, processamento das ordens de trabalho e informação fiabilística obtida. Os protótipos serão explicados em termos de funcionalidade, processo e variáveis de entrada e de saída, sendo a parte do código desenvolvida em anexo.

Numa parte introdutória será apresentado o *software* utilizado na realização do projeto e as razões pelas quais foi escolhido. De seguida é feita uma descrição da fase de importação de dados e do pré-processamento, anterior à inserção dos dados nos modelos de previsão. Será feita também a comparação dos modelos e das técnicas de pré-processamento descritos no enquadramento teórico deste projeto, recorrendo às métricas de avaliação de desempenho. Será desenvolvida a utilização destes modelos para a previsão de tipo de ordem de trabalho e do modo de falha de um equipamento, e quais os modelos a adotar nestas soluções.

Numa fase final será descrita a implementação da análise fiabilística das ordens de trabalho processadas, com a comparação dos dois modos de parametrização estudados, com a criação das funções necessárias para a comparação de distribuições e elações que se podem retirar a partir dos dados.

4.1 *Software* Utilizado

Para uma abordagem de protótipo, e devido à sua facilidade de interpretação, a linguagem de programação utilizada foi o *Python*, uma linguagem de alto nível. Outras vantagens desta linguagem baseiam-se na vasta quantidade de bibliotecas, que oferecem soluções mantidas e atualizadas, de modo a reduzir a carga de programação ao utilizador, facilidade na definição de variáveis, classes e funções e, em regra geral, flexibilidade ao código desenvolvido e impermeabilidade de erros.

A distribuição utilizada foi a *Anaconda*, constituída por um aglomerado de distribuições indicadas para análise de dados e *machine learning*, e um navegador para soluções de interface com o utilizador (*Jupyter Notebook*). Este *software* é recomendado também para a visualização de resultados, sendo o código facilmente corrido, e verificando-se alterações no output com a variação do código. Além disso, o código pode ser separado em células, onde apenas uma parte é corrida, ao contrário de outras alternativas que interpretam ou correm um ficheiro todo quando este é chamado.

O IDE (Ambiente de desenvolvimento integrado) utilizado foi o *Pycharm*. Este *software* permite a criação de módulos de código *python*, que serão conjuntos de classes e funções mais sistematizadas e escaláveis. Apresenta também uma solução de inspeção de código, capaz de corrigir erros possíveis ou más práticas de programação.

4.2 Importação dos Dados

A primeira fase da análise de dados passa sempre pela importação dos dados existentes para o programa utilizado. Neste caso, os dados existentes apresentam-se de três maneiras diferentes; formato de ficheiro de texto, formato de ficheiro excel e formato de *comma-separated values*.

A especificação fornecida pela EQS contém os parâmetros possíveis a retirar de uma ordem de trabalho, de modo a serem inseridos numa tabela e posteriormente numa base de dados. No entanto, nem todos os parâmetros estão disponíveis nos dados fornecidos, visto que a especificação é abrangente e pretende uma análise mais detalhada da que é feita no momento.

A importação dos dados foi feita com recurso à libraria *Pandas* (McKinney, 2011), que é responsável por importar, exportar e alterar dados em forma de tabela (denominada de *Dataframe*) ou vetor indexado (denominado de *Series*).

Para importação foram utilizados o primeiro conjunto de dados apresentados, pois estes são os que se encontram menos tratados para abordagens seguintes, e o segundo conjunto, pois vem em formato excel com varias folhas e é o formato mais comum de dados. O terceiro conjunto de dados tem uma importação direta com a utilização de um método existente na biblioteca *Pandas*.

No primeiro caso, para contornar o facto da utilização de um ficheiro em forma de relatório, este foi transformado em *comma-separated values*. O símbolo de segmentação de colunas escolhido foi o da barra vertical "|" (o escolhido por defeito é a virgula "," pois os ficheiros normalmente apresentam a vírgula como o separador e daí a origem da extensão do ficheiro).

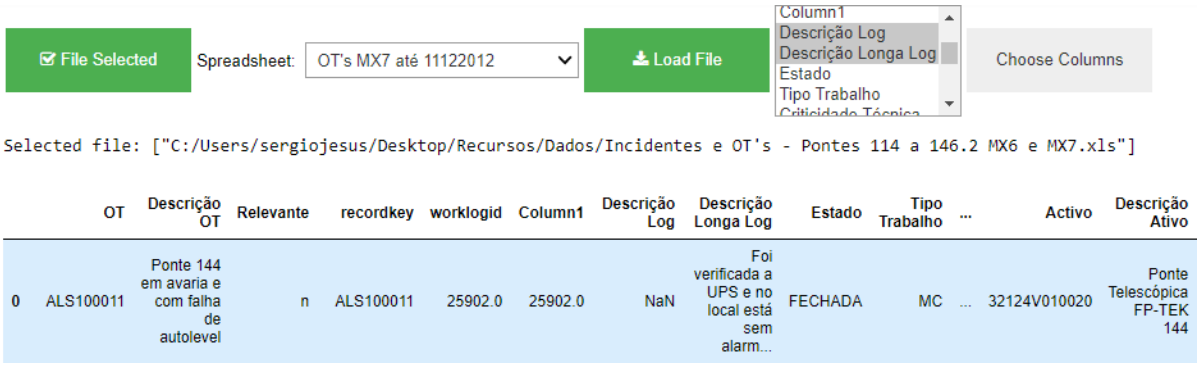
Na parte de processamento, foram removidas colunas inexistentes que surgiram devido à formatação dos dados e foram removidos todos os espaços brancos que existiam no ficheiro original. As colunas foram renomeadas para as designações fornecidas pela especificação e este ficheiro foi exportado com um formato mais utilizável por um *software* de desenvolvimento.

	wo_type	wo_finish_datetime	wo_no	eq_id	op_unit_desc	wo_header_short	wo_priority	wo_plan_cost	wo_total_cost	wo_status
0	COR1		100000015296	45200000100001	ASTILLADORA	sus. placa desgaste cuchilla	2	0.00	2,097.50	
1	COR1		100000016151		ASTILLADORA	sustitución correas astilladora	2	0.00	9,861.80	
2	COR1		100000017011		ASTILLADORA	sus rodillo pinchos astilladora	2	0.00	2,248.00	
3	COR1		100000018268		ASTILLADORA	limpieza mandibula astilladora	2	0.00	502.52	
4	COR1		100000025435		ASTILLADORA	sus. correas astilladora	2	0.00	10,141.57	
5	COR1		100000025840		ASTILLADORA	sus. rodillo pinchos alimet astilladora	5	0.00	5,252.23	

Figura 10 - Formato dos dados após tratamento

Como é possível observar, existem campos que foram preenchidos em apenas parte de todas linhas, e por essa razão, permanecem vazios nas restantes.

No segundo caso, foi desenvolvido um protótipo de GUI (Interface gráfica para o utilizador), constituído com menus de seleção e botões, capazes de importar os dados de um formato excel e selecionar as colunas relevantes, onde se encontram os campos de texto.



	OT	Descrição OT	Relevante	recordkey	worklogid	Column1	Descrição Log	Descrição Longa Log	Estado	Tipo Trabalho	...	Activo	Descrição Ativo
0	ALS100011	Ponte 144 em avaria e com falha de autolevel	n	ALS100011	25902.0	25902.0	NaN	Foi verificada a UPS e no local está sem alarm...	FECHADA	MC	...	32124V010020	Ponte Telescópica FP-TEK 144

Figura 11 - Protótipo de GUI, com recurso a *Widgets* do *Jupyter notebook*

Esta solução permite a escolha do ficheiro, com o recurso a um navegador de ficheiros, quando selecionado o primeiro botão, escolher a folha de cálculo relevante para a importação com o *Dropdown Menu*, e escolha das colunas onde se encontram os documentos para a formação de uma variável de documentos em forma de *Series*, passível de processamento posterior.

4.3 Soluções de Análise de Texto e Extração dos Dados

Após a extração dos documentos em forma de texto, seguiu-se um tratamento dos dados. O primeiro passo foi a criação de um *Bag of Words*, um aglomerado de todas as palavras existentes nos documentos, para que sejam postos em prática os algoritmos de TF-IDF e de CBTW mencionados anteriormente.

4.3.1 Tokenização

O processo de *tokenização* utilizado baseou-se numa biblioteca existente de tratamento de texto, *SpaCy*, e na utilização de *Regular Expressions* para refinar o processo e otimizá-lo para o caso da língua portuguesa.

A biblioteca *SpaCy* disponibiliza soluções de análise e tratamento de texto eficaz com capacidade de processamento elevada, sendo possível a utilização para análise de elevados números de documentos. Como alternativa, poderiam ter sido utilizadas soluções existentes na biblioteca de *Natural Language ToolKit*, ou uma solução exclusiva de *Regular Expressions*. A primeira foi evitada pois apresenta soluções de fácil compreensão, mas de pouco eficientes, sendo o processamento mais lento que o da biblioteca *SpaCy* (Honnibal e Johnson, 2015), enquanto que a solução por *Regular Expressions* é mais permeável a erros e ignora resultados que tenham caracteres não contemplados pelo utilizador, apesar de apresentar uma eficiência muito maior dos casos anteriores.

As *Regular Expressions* foram posteriormente desenvolvidas utilizadas para a remoção de carateres que passam no teste dos métodos da libraria *SpaCy* não existentes na língua portuguesa, e mantendo os existentes. Além disso, foram utilizadas para a separação de palavras ligadas por pontuação que não o travessão; palavras ligadas por pontos finais, virgulas, números e outros carateres eram reconhecidos erradamente como *tokens* antes desta fase de processamento.

A Figura 12 serve como exemplo ao processamento feito pela função de tokenização desenvolvida. É de notar que tem não só como função a separação individual das palavras, mas também normalização do texto, retirando acentuação e maiúsculas.

```
Input: Avaria no conjunto motor-bomba; Fuga de óleo por desgaste da bucha.
Output: ['avaria', 'no', 'conjunto', 'motor-bomba', 'fuga', 'de', 'oleo', 'por', 'desgaste', 'da', 'bucha']
```

Figura 12 - Exemplo de Tokenização de um documento

Devido à inexistência de soluções para a língua portuguesa, e pela complexidade acrescente, optou-se por não se realizar truncatura ou lematização das palavras, deixando o resultado da tokenização como input para tratamento posterior.

4.3.2 Correção de Erros Ortográficos

Ao ler os dados existentes, é notória a existência de erros de ortografia e abreviaturas que para um leitor são facilmente detetáveis e corrigíveis, mas que passam no processo de tokenização desenvolvido como sendo palavras diferentes, e consequentemente, para os algoritmos posteriores de TF-IDF e CBTW, significados diferentes.

Dicionário por Web Crawler

Para solucionar este problema, foi necessária a criação de um ficheiro com as palavras constituintes do vocábulo português, para comparação com os *tokens* resultantes da fase anterior.

Este foi conseguido pela utilização de técnicas de *Web Crawling*, que consiste num algoritmo que percorre automaticamente páginas na internet e retira a informação pretendida. Como nem todos os *websites* permitem que seja feita uma extração de dados, devido às suas políticas de direitos de autor e de privacidade, teve que ser encontrado um site permissivo para ser acedido múltiplas vezes, e que permitisse a utilização da informação contida.

Após escolhido o *website* que melhor se ajustou ao pretendido, foi desenvolvido um algoritmo capaz de realizar a análise sintática do HTML da página, recorrendo à libraria *bs4* para a apresentação dos dados de uma maneira mais compreensível ao utilizador.

Após ser encontrado o padrão de formulação de uma página, o algoritmo foi corrido iterativamente até serem esgotados os registos de palavras novas a serem pesquisadas no *website*, sendo a cada iteração produzida uma página diferente. A cada iteração, a página da palavra era lida, as palavras de toda a página eram retiradas e as novas palavras que ainda não se encontravam no dicionário eram registadas. Além disso, um segundo campo era preenchido, de modo a identificar que a palavra tinha sido pesquisada.

Como resultado final, obteve-se um dicionário com cerca de 63 mil entradas, obtidas num intervalo de 8 horas. Este intervalo foi extenso visto que cerca de 90% do tempo (1 segundo de paragem por cada página lida), o algoritmo encontrava-se suspenso, de modo a não interferir com o funcionamento normal dos servidores envolvidos.

As palavras resultantes foram normalizadas (removendo-se pontuação e maiúsculas), reduzindo em cerca de mil entradas no número final de palavras.

Remoção dos erros

Com o dicionário criado, foi possível a comparação com os *tokens* existentes nos documentos das ordens de trabalho. Inicialmente, todas os *tokens* foram cruzados, e todos os que foram encontrados no dicionário foram assumidos que estavam escritos corretamente. Posteriormente, os *tokens* mais frequentes (que ocorressem mais de 100 vezes na totalidade dos documentos) foram também considerados corretos, visto que apenas a frequência é tida em conta nos algoritmos seguintes e *tokens* mais frequentes são menos importantes para o algoritmo. Por fim, os *tokens* que não se encaixavam em nenhuma das duas categorias anteriores foram comparados com tokens já existentes pelo método da distância de edição.

O cálculo da distância de edição de duas palavras consiste no número mínimo de movimentos necessários para transformar uma palavra em outra. Os movimentos contabilizados são a adição, subtração e substituição de caracteres.

Tabela 4 - Operações visadas na distância mínima de edição (exemplos retirados dos dados estudados)

Operação	Palavra Inicial	Palavra Final
Adição	Fol	Fole
Subtração	Correcção	Correção
Substituição	Acaria	Avaria

Cada uma destas operações tem como distância de edição o valor de 1. A distância de edição de Fol para Fole tem o valor de 1, enquanto que a distancia de *Levenshtein* de Fol para Folha tem o valor de 2, pois necessita de duas operações de adição. Isto significa que a correção adotada será a primeira, pois é a que apresenta o menor valor de distância de edição. O valor máximo de distância de edição para substituição adotado foi de 2, pois a partir deste valor, as palavras apresentam diferenças significativas. O critério de desempate a partir de duas palavras com a mesma distância de edição escolhido foi o de frequência máxima, ou seja, a palavra escolhida é a que mais vezes aparece na totalidade dos documentos.

Foi então calculada uma matriz com as distâncias entre as palavras não encontradas no dicionário e as palavras encontradas. Caso se utilizasse novamente o dicionário para a criação da matriz, esta teria dimensões muito maiores e não traria nenhum benefício, pois os algoritmos de TF-IDF e de CBTW são "cegos" a erros, ou seja, uma palavra escrita corretamente que apareça uma vez tem o mesmo efeito que uma palavra escrita incorretamente que apareça uma vez. Mesmo com esta simplificação, o processamento é demorado, e a matriz resultante tem cerca de 26 milhões de elementos. A matriz foi reduzida a um vetor de mínimos, no qual se encontrava os índices para a substituição das palavras escritas incorretamente. Palavras com distância de edição superior a 2 foram consideradas corretas também, para operações posteriores.

4.3.3 Input dos Modelos para Comparação

A implementação do TF-IDF foi feita com base nos métodos da biblioteca *Scikit-learn* (Pedregosa et al., 2011), que apresenta soluções apropriadas a documentos na forma de *Dataframe* ou *Series* da biblioteca *Pandas*, permitindo uma adaptação rápida ao formato dos dados.

Estes métodos permitem a alteração de parâmetros anteriormente referidos, como é o caso de número de N-gramas, alteração dos parâmetros de suavização ou da função de normalização.

O output deste método é uma matriz esparsa, com o número de documentos como número de linhas e o número de *tokens* existentes como número de colunas (matriz com dimensões de 8.000×10.000), com um preenchimento de 0,3% de todos os elementos da matriz (cerca de 22.000 elementos). Este valor é explicado pelo facto de apenas parte pequena de todas as palavras utilizadas é utilizada em cada documento, o que significa que por cada linha irá existir uma quantidade reduzida de elementos.

Para *input* dos modelos seguintes, o cálculo do TF-IDF seguiu-se sem a utilização de N-gramas (os *tokens* não foram agrupados), visto que estes introduziam maior número de parâmetros de entrada, e consequentemente maiores tempos de processamento para os modelos. Utilizaram-se também os parâmetros de suavização e normalização genéricos do método (normalização logarítmica e suavização adicionando uma unidade a cada membro da divisão).

4.4 Comparação dos modelos de *Machine Learning*

Para a comparação e escolha do melhor modelo de *machine learning*, foi utilizado o método de validação cruzada com $k = 10$, recorrendo ao TF-IDF calculado a partir do segundo conjunto de dados anterior ao pré-processamento; para manter a integridade do teste, todos os modelos receberam o mesmo input, e tinham como objetivo determinar se uma ordem de trabalho era do tipo corretivo ou preventivo pelas descrições existentes.

Para garantir que o input de cada modelo é igual, foi colocada uma *seed* para que a geração de números aleatórios produzisse resultados constantes na função de divisão em amostras, chamada de *Kfold*.

Os modelos simples utilizados para comparação foram:

- Árvores de Decisão
- *Naïve Bayes*
- Máquina de vetores Suporte (linear)
- Regressão Logística
- Redes Neuronais Artificiais (com 5 camadas)

A comparação dos modelos foi feita com a utilização das métricas de *F measure* e *ROC-AUC*.

No caso deste conjunto de dados, confirma-se o desequilíbrio previsto, sendo a classe maioritária a de ordens de trabalho corretivas, perfazendo 97% dos resultados, enquanto que a classe minoritária é a de ordens de trabalho preventivas, que perfazem os restantes 3%.

Os resultados obtidos são mostrados na tabela 5:

Tabela 5 - Pontuações dos modelos simples para classificação

Modelo	F measure	ROC-AUC
Naïve Bayes	0,70	0,98
Árvores de Decisão	0,91	0,95
SVM	0,88	0,99
Regressão Logística	0,84	≈1,00
ANN	0,89	0,98

Com a análise da *F measure* é possível concluir que, para este caso, o modelo com melhor desempenho é o de Árvores de Decisão, seguido pelas redes neuronais artificiais e máquinas de vetor suporte. O modelo mais simples, *Naïve Bayes*, obteve um resultado muito baixo em relação aos restantes modelos, visto que não é recomendável para dados desequilibrados; tende a classificar as amostras com a classe dominante.

Os valores de ROC-AUC obtidos foram todos próximos do valor máximo, não podendo assim retirar qualquer conclusão com esta métrica.

Para observar a alteração provocada pelo aumento do número de camadas de neurónios numa rede neuronal artificial, e consequentemente, aumento a complexidade do modelo, foi replicado o mesmo teste com a variação do número de camadas.

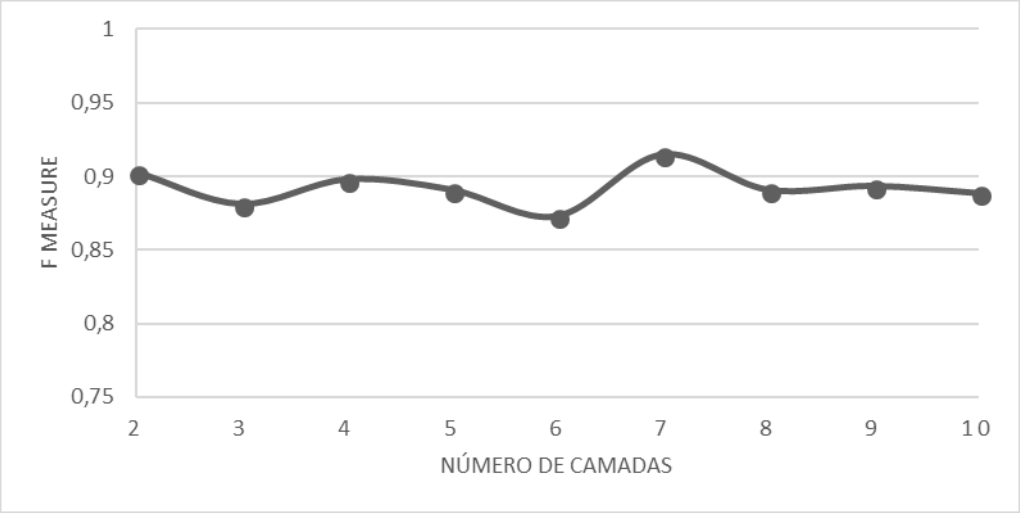


Figura 13 - Gráfico da evolução de *F measure* com o número de camadas de neurónios

Analisando o gráfico obtido, o valor do medidor de desempenho do modelo permanece não sofre variações muito significativas, tendo um máximo com 7 camadas e um *F measure* de 0,91 e um mínimo com 6 camadas e um *F measure* de 0,87. Isto significa que o aumento de complexidade do modelo, que por consequência aumenta a capacidade de interpretação dos dados não aumenta significativamente a performance do modelo; a otimização do processo deverá ser tratada ao nível do pré-processamento dos dados.

Os modelos analisados anteriormente estavam disponibilizados na biblioteca de *machine learning* para *Python*, *SkLearn*. No entanto, como o modelo NBSVM se trata de um modelo composto, e são utilizados elementos dos dois modelos constituintes para o resultado final, um algoritmo teve que ser desenvolvido para a sua utilização. Com a passagem do modelo SVM para um NBSVM foi possível a passagem de um *F measure* de 0,88 para 0,90. Isto vem a confirmar que a utilização de dois modelos simples em cadeia produz um resultado melhor que os modelos individuais.

O valor dos indicadores é, por norma, próximo da realidade, tendo alterações em amostras não classificadas a partir da segunda casa decimal. Esta característica foi verificada com os resultados da participação na competição inicial de processamento de linguagem natural.

4.5 Impacto de Modificações no Pré-processamento nos Modelos

Para esta comparação, foi utilizado o modelo de NBSVM, fazendo variar o pré-processamento. Utilizaram-se três tipos diferentes de pré-processamento; o método já anteriormente referido de CBTW, o método de TF-IDF e o método de matriz de frequências normalizadas.

Como se trata de um caso de grande desequilibro das classes, decidiu-se utilizar a métrica de CBTW, que contabiliza a proporção da classe com menor número de registos e atribui aos parâmetros existentes nas amostras que a constituem um maior peso.

A implementação do CBTW e da matriz de frequências normalizadas foi feita a partir da criação de um módulo, com *inputs* e *outputs* de tipos de variáveis iguais ao TF-IDF da libraria *SkLearn*, para que seja fluída a passagem de um para outro tipo de pré-processamento na fase de cálculo das métricas de avaliação.

O valor das métricas foi calculado para dez ensaios de diferentes divisões em *Kfold*, aumentando artificialmente o volume de amostras e a segurança do valor de *F measure*.

Tabela 6 - Comparação da performance de diferentes pré-processamentos

Tipo de Pré-Processamento	F measure
TF-IDF	0.8923
NTF	0.9048
CBTW	0.9341

Naturalmente, a performance do método de CBTW no *F measure* é superior aos restantes, devido ao impacto da classe minoritária no *input* do modelo e a medida *F measure* privilegear a classificação acertada de classes minoritárias. No entanto, o método de matriz de frequências normalizada apresentou ligeiramente melhores resultados aos do TF-IDF, apesar de conter mais informação.

Este resultado pode dever-se ao facto de muitos dos documentos provenientes das ordens de trabalho se encontram duplicados, e com a ponderação da frequência inversa nos documentos, os termos utilizados que poderão ser relevantes são tratados como palavras vazias e têm um peso menor no modelo.

A outra diferença possível de ser realizada durante o pré-processamento é a correção dos erros de escrita, com recurso ao dicionário e ao método da distância de edição mínima referido anteriormente. As imagens 15 e 16 mostra alguns exemplos do algoritmo de correção automática de erros, onde do lado esquerdo se apresentam as palavras originais e do direito se apresentam as correções.

```
['alame', 'alarme'],
['alames', 'alarmes'],
['alamre', 'alarme'],
['alamres', 'alarmes'],
['alare', 'alarme'],
['alarem', 'alarme'],
['alareme', 'alarme'],
['alarems', 'alarme'],
['alarm', 'alarme'],
['alarmea', 'alarme'],
['alarmme', 'alarme'],
['alarmne', 'alarme'],
['alarne', 'alarme'],
['alarrme', 'alarme'],
['alarrmes', 'alarmes'],
```

Figura 14 - Exemplo de correções acertadas do algoritmo de correção

Todavia, visto que o dicionário criado por *web scrapping* não contem todas as palavras portuguesas e todos os tempos verbais, e também por limitações no algoritmo de cálculo de distância de edição algumas palavras foram existentes corrigidas erradamente.

```
['alertadas', 'apertadas'],
['alertado', 'apertado'],
['alertados', 'apertado'],
['alertamos', 'aceitamos'],
['alertando', 'apertado'],
```

Figura 15 - Exemplos de correções erradas do algoritmo de correção

Para a avaliação das alterações provocadas com este pré-processamento nos documentos, foi escolhido novamente fazer dois testes para texto não corrigido e texto corrigido com dez ensaios cada. O pré-processamento escolhido a partir dos documentos foi o TF-IDF e o modelo de avaliação foi o NBSVM.

Tabela 7 - Comparação da remoção de erros no texto

Remoção de erros	F measure
Sim	0.8807
Não	0.8923

4.6 Decisão do Modelo e de Pré-processamento Realizado

Para o caso de análise de texto de ordens de trabalho, de modo a ser classificado o tipo da mesma, o modelo ideal mostrou ser o NBSVM, com o método de pré-processamento CBTW e sem a remoção de erros de escrita.

Este resultado vem ao encontro à pesquisa realizada por Wang e Manning (2012) e de Liu et al. (2007), que mostraram que para a classificação de sentimento, o modelo de NBSVM é capaz de obter resultados melhores que modelos mais complexos (como é o caso das redes neuronais artificiais), e que para casos de desequilíbrio entre classes, o CBTW é uma alternativa mais robusta que o TF-IDF.

O processo de remoção de erros provocou uma ligeira queda no *F measure* do modelo, e as razões possíveis para isto acontecer poderão passar pela diminuição da dimensionalidade dos dados e consequente perda de informação, e também das limitações referidas anteriormente, pela correção errada de palavras.

Estes algoritmos poderiam ser escalados para o caso de classificação do modo de falha de um equipamento, mas para isso seria necessário que os dados dos dois primeiros conjuntos já tivessem uma infraestrutura preparada para este tipo de aplicação. Todavia, estes mesmos algoritmos poderão ser aplicados, caso numa amostra sejam classificados manualmente os modos de falha, e a previsão poderá ser alargada para o resto dos dados.

4.7 Análise Fiabilística

Após do pré-processamento e escolha dos métodos de *machine learning* para a análise de texto, seguiu-se a automatização de técnicas de análise fiabilística.

4.7.1 Tratamento dos Dados

Para a análise fiabilística foi necessário inicialmente tratar os dados, visto que não estavam calculados os intervalos de tempo entre avaria ou substituição nem separados entre tempos completos e censurados.

	datetime	machineID	comp	IF_FAIL
0	2014-07-01 06:00:00	1	comp4	0
1	2014-09-14 06:00:00	1	comp1	0
2	2014-09-14 06:00:00	1	comp2	0
3	2014-11-13 06:00:00	1	comp3	0
4	2015-01-05 06:00:00	1	comp1	0
5	2015-01-20 06:00:00	1	comp1	0
6	2015-02-04 06:00:00	1	comp3	1
7	2015-02-19 06:00:00	1	comp3	0
8	2015-03-06 06:00:00	1	comp3	0

Figura 16 - Dados antes do tratamento

A melhor maneira de apresentação dos dados para aplicação em algoritmos posteriores encontrada foi a forma de tabela de valores (*DataFrame*), separados entre tempos completos e tempos censurados, separados por ID da máquina em linha e por componente em coluna.

Foi desenvolvido um módulo deste processamento, de modo a que procedimentos semelhantes possam ser adaptados a esta forma de apresentação de dados.

	comp1	comp2
1	[105, 113, 15, 60, 15, 15, 210, 46]	[105, 248, 30, 15, 15, 90, 76]
2	[30, 245, 45, 15, 30, 15, 60, 45, 30, 64]	[165, 80, 60, 60, 45, 45, 15, 109]
3	[0, 220, 75, 15, 255, 14]	[0, 265, 15, 60, 45, 15, 150, 15, 14]
4	[0, 269, 60, 105, 120, 25]	[75, 179, 45, 45, 30, 45, 30, 90, 40]
5	[60, 189, 15, 30, 60, 45, 30, 45, 15, 30, 30, 30]	[15, 204, 15, 75, 15, 15, 15, 30, 15, 30, 15, ...
6	[180, 72, 105, 75, 45, 45, 15, 42]	[165, 177, 15, 90, 15, 45, 72]
7	[105, 140, 75, 30, 30, 45, 15, 15, 124]	[150, 80, 30, 75, 30, 30, 15, 45, 15, 15, 30, 64]

Figura 17 - Dados após o tratamento

4.7.2 Módulo de Distribuições

Um módulo de distribuições anterior ao projeto já tinha sido desenvolvido por elementos da equipa da EQS. Este módulo é capaz de ajustar os parâmetros de diversas distribuições, para tempos completos. No entanto, funcionalidades adicionais foram criadas, de modo a que este módulo se ajustasse melhor ao pretendido.

A primeira alteração realizada foi a da criação de um teste de Laplace, para observar a tendência dos dados. Para garantir que o algoritmo estava a funcionar normalmente, foi utilizado um $\alpha = 5\%$ e o teste foi corrido para dez mil amostras de tamanho 4 de uma distribuição de *Weibull* de parâmetros $\alpha = 200$ e $\beta = 1{,}5$ sendo α o fator de escala e β o parâmetro de forma.

Tabela 8 - Resultados da simulação de Teste de Laplace

Resultado	Frequência	Proporção
Tempos IID	9950	99,5%
Tendência Crescente	28	0,28%
Tendência Decrescente	22	0,22%

Com os resultados simulados é possível detetar que existe apenas um valor residual de amostras classificadas erradamente, e na verdade, o valor do erro do tipo α no teste de hipóteses é muito mais baixo que os 5% estipulados (para o caso da distribuição de *Weibull* definida).

A segunda alteração foi a possibilidade de introdução de tempos censurados para a estimação de parâmetros.

O método de estimação de parâmetros está estruturado para a criação de uma função de verosimilhança, com o cálculo da função de densidade de probabilidade dos pontos e posterior multiplicação; a otimização dá-se calculando o máximo desta função, com métodos numéricos iterativos implementados na biblioteca *SciPy*. A ausência de uma parcela da função de probabilidade cumulada significa que este algoritmo não contempla tempos censurados.

Este objetivo foi conseguido com a introdução do cálculo da função de probabilidade cumulada dos dados censurados, e multiplicando a parcela à função de verosimilhança.

4.8 Comparação entre LSM e MLE

Para definir a metodologia a adotar entre o método de mínimos quadrados e o de máxima verosimilhança, foi necessário comparar os dois métodos. Ambos os métodos foram testados sem a utilização de dados censurados, para uma previsão mais próxima à real.

Foram geradas dez amostras de tamanho dez, aleatoriamente, de uma distribuição de *Weibull*. Para cada uma das amostras, foram utilizados os dois métodos para a estimação dos parâmetros da distribuição de *Weibull*.

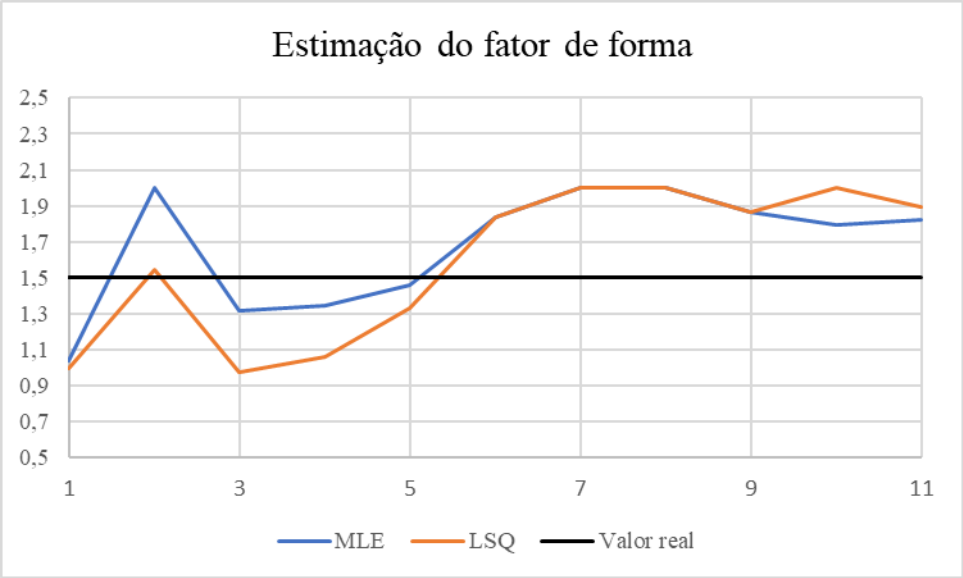


Figura 18 - Gráfico dos resultados da estimação do parâmetro fator de forma

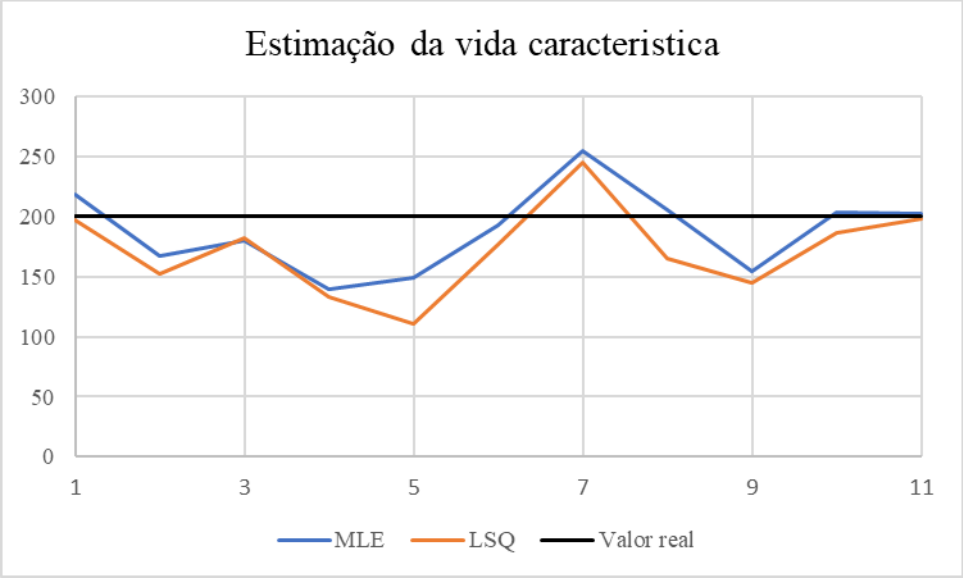


Figura 19 - Gráfico dos resultados da estimação da vida característica

O erro médio dos dois estimadores está representado na tabela 9.

Tabela 9 - Erro médio de estimação

Resultado	α	β
MLE	0.120	11.469
LSQ	0.138	14.789

O método dos mínimos quadrados obteve consistentemente valores mais próximos dos reais da distribuição subjacente.

4.8.1 Parâmetros de Análise Calculados

A partir do momento em que se obtiveram os parâmetros da distribuição associada ao modo de falha, com recurso à função da distribuição, foi possível calcular os seguintes valores ou funções, para a análise posterior:

Tabela 10 – Parâmetros de manutenção calculados

Parâmetro	Definição	Fórmula
Função densidade de probabilidade de falha	Proporção de falhas ocorridas no instante t (função parametrizada anteriormente)	$f(t)$
Função de probabilidade acumulada de falha	Proporção de falhas ocorridas até ao instante t	$S(t) = \int f(t)dt$
Função de risco	Probabilidade instantânea de um equipamento se avariar	$h(t) = \frac{f(t)}{S(t)}$
Fiabilidade	Probabilidade de um equipamento não avariar até ao instante t	$R(c) = 1 - S(t)$
MTTF	Tempo médio de falha de um componente	$MTTF = \int t \times f(t)dt$
$\boldsymbol{\lambda}$	Taxa de falhas	$\lambda(T) = \frac{d}{dT}E[N(T)]$
MTBF	Tempo médio de falha de um equipamento	$MTBF = \frac{1}{\lambda}$
Probabilidade de falha num intervalo de tempo	Probabilidade de ocorrência de uma falha entre o tempo t_1 e t_2	$\mathrm{P} = \int_{t_1}^{t_2} f(t)dt$

5 Considerações Finais e Perspetivas de Trabalhos Futuros

Neste capítulo irão ser feitas reflexões e elações retiradas no decorrer deste projeto em relação aos tópicos abordados. Posteriormente, serão apresentados novos caminhos de desenvolvimento descobertos e temas não tratados que poderão vir a ser motivo de análise mais detalhada futuramente.

5.1 Considerações Finais

Com este projeto foi possível avaliar as potencialidades de utilização de modelos de *Machine Learning* aplicados à análise de texto das ordens de trabalho. Esta utilização torna-se mais relevante quando se trata de dados de dimensões industriais, onde a utilização de mão-de-obra humana provoca uma análise pouco eficiente, demorada e incompleta.

Foi possível neste estudo a perceção das diferenças entre cada modelo, na sua génese, modo de operação, limitações, aplicabilidade e complexidade. A utilização de diferentes métodos para a criação de parâmetros para os modelos (o pré-processamento dos dados) comprovou ser um dos aspetos mais importantes, onde é possível aplicar soluções mais criativas, de modo a tornar obvio os padrões nos dados para a solução desejada. Muitas destas soluções podem ser adaptadas a outros casos de análise de sentimento e inclusivamente otimizadas, como é o caso dos módulos de CBTW, os métodos de correção de erros ortográficos e as métricas de avaliação dos modelos. A utilização de ferramentas existentes na comunidade do *python* mostrou ser crucial, visto que soluções que não seriam facilmente implementadas, como é o caso dos modelos apresentados, já se encontram programados de uma forma generalizada e pronta para o utilizador.

Quanto à análise fiabilística, uma solução programada e corretamente otimizada permite o processamento de dados de uma forma mais eficiente em comparação a soluções tradicionais, podendo também serem feitas análises a milhões de dados de uma forma rápida e à prova de erros de transcrição. No entanto, na realização do trabalho, foi possível notar que, pelo menos nos casos estudados, as ordens de trabalho poderão sofrer melhorias de modo a adaptarem-se a uma realidade em que toda a informação pode ser processada, tratada e analisada de uma forma quase instantânea. Esta melhoria poderia, por exemplo, ser feita ao nível do documento em si, com o registo de mais informação e especificar campos para cada equipamento, ou ao nível do registo, habilitando o técnico responsável pela manutenção do equipamento o registo em tempo real dos sintomas e das operações realizadas.

5.2 Perspetivas de Trabalhos Futuros

Para desenvolvimento futuro, a fusão de um modo mais intrínseco destas duas áreas de conhecimento (Ciência de Computação e Manutenção) parece ter grande potencialidade, aproveitando não só a descrição de ordens de trabalho recentes, mas também dados de sensorização de equipamentos, alarmes e outra informação relevante para o processo, como os dados que se encontram no terceiro *dataset*.

A área da inteligência artificial está em desenvolvimento constante, e o aparecimento de novas técnicas com melhores resultados ocorre com uma frequência elevada, neste momento. Para este projeto foram utilizados modelos mais básicos, para a sua melhor perceção conceitualmente e utilização. Além disso, neste projeto apenas foram usados modelos supervisionados; para problemas em que não existe uma amostra de treino a utilização de modelos não supervisionados, como os de *clustering* (*K-means, ANN, Expectation-Maximization),* é uma solução viável.

Para um estudo mais completo da eficiência modelos aplicados, a utilização de mais e melhores métricas, como a perda logarítmica (*Log-Loss*), ou a interpretação da matriz de confusão completa.

Quanto à análise fiabilística, a integração de procedimentos e ferramentas mais sofisticados de análise de múltiplos modos de falha (como o FMEA e FMECA) aparenta ser um caminho a desenvolver. A utilização de testes de hipótese do ajuste como o teste Qui-Quadrado, Kolgorov-Smirnov ou o critério de Cramér–von Mises parecem ser avaliações mais completas dos resultados. Além disso, a contemplação de soluções para casos de modos de falha dependentes deve ser abordada. A existência de campos de custo de substituição preventiva e corretiva, num *Dataset*, poderão levar a uma análise de tempos de substituição ideais, impossibilitada pela natureza dos dados utilizados neste projeto.

Referências

Assis, Rui. 2004. "Apoio à Decisão Em Gestão Da Manutenção", Lidel - Edições Técnicas, Ldª. pp-48-51.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. *Pattern Recognition*. Vol. 4. https://doi.org/10.1117/1.2819119.

Bolch, Gunter, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi. 2006. "Queueing Networks and Markov Chains." https://doi.org/10.1002/0471791571.

Bramer, Max. 2016. *Principles of Data Mining*. *Springer*. https://doi.org/10.1007/978-1-4471-7307-6.

Cabral, José Paulo Saraiva. 2006. "Organização e Gestão Da Manutenção", Lidel - Edições Técnicas, Ldª.pp.93-118.

Hand, David J., and Robert J. Till. 2001. "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems." *Machine Learning* 45 (2): 171–86. https://doi.org/10.1023/A:1010920819831.

Honnibal, Matthew, and Mark Johnson. 2015. "An Improved Non-Monotonic Transition System for Dependency Parsing." *Emnlp 2015*, no. September: 1373–78. http://aclweb.org/anthology/D15-1162.

Jurafsky, Daniel, and James H Martin. 2017. "Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition." https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Lee, Elisa T., and John Wenyu Wang. 2003. *Statistical Methods for Survival Data Analysis*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471458546.

Liu, Ying, Han Tong Loh, Youcef Toumi Kamal, and Shu Beng Tor. 2007. "Handling of Imbalanced Data in Text Classification: Category-Based Term Weights." In *Natural Language Processing and Text Mining*, 171–92. https://doi.org/10.1007/978-1-84628-754-1_10.

McKinney, Wes. 2011. "Pandas: A Foundational Python Library for Data Analysis and Statistics." *Python for High Performance and Scientific Computing*, 1–9.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "[Scikit-Learn] Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30. https://doi.org/10.1007/s13398-014-0173-7.2.

Piantadosi, Steven T. 2014. "Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions." *Psychonomic Bulletin & Review* 21 (5): 1112–30. https://doi.org/10.3758/s13423-014-0585-6.

Runkler, Thomas A. 2016. *Data Analytics. Models and Algorithms for Intelligent Data*

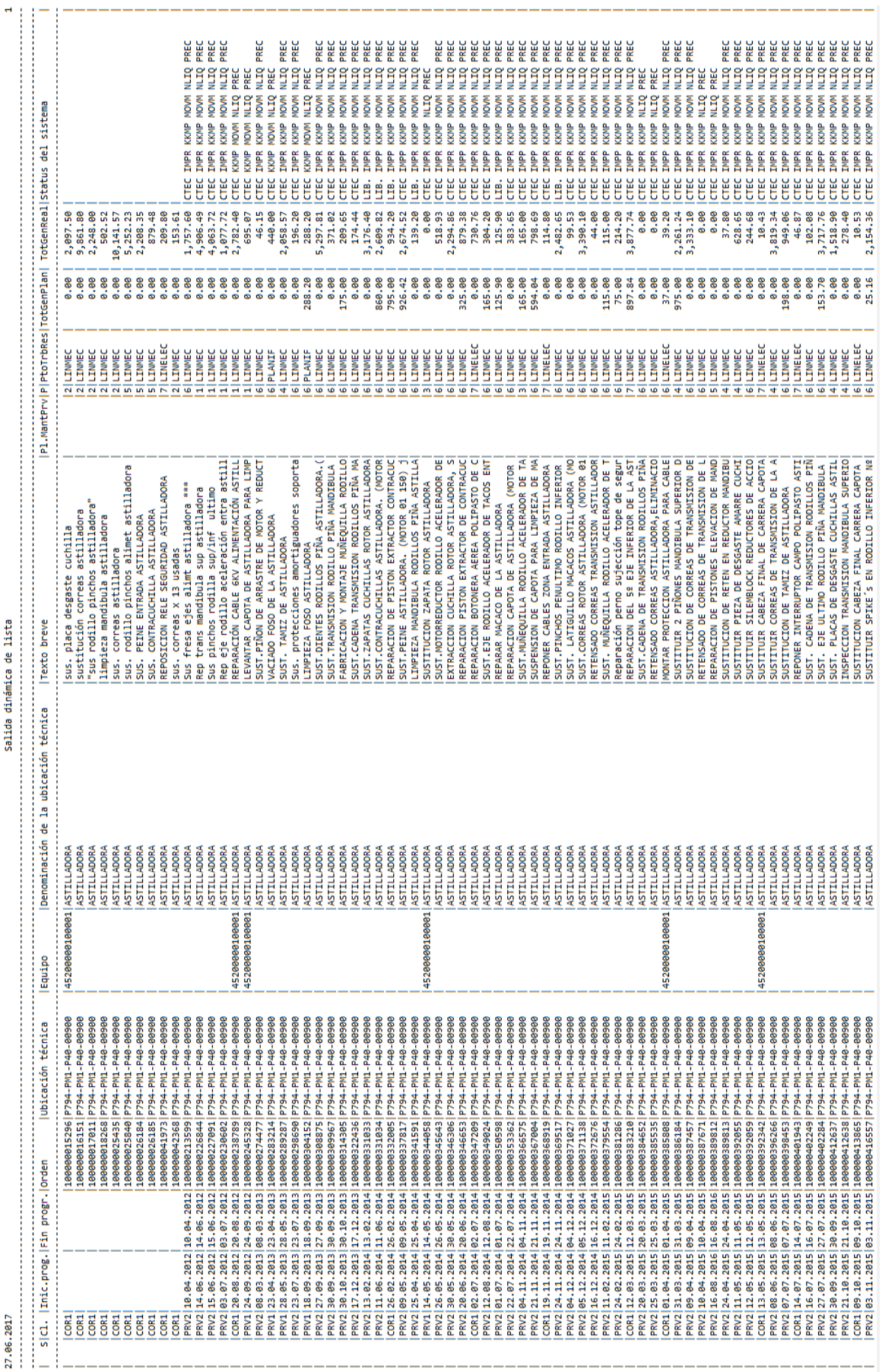
Analysis. *Vasa*. https://doi.org/10.1007/978-3-8348-2589-6.

Tirunagari, Santosh. 2015. “Data Mining of Causal Relations from Text: Analysing Maritime Accident Investigation Reports.” *Working Paper*. http://arxiv.org/abs/1507.02447.

Wang, S, and C Manning. 2012. “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification.” *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, no. July: 90–94.

ANEXO A:*Datasets* utilizados

1º Dataset



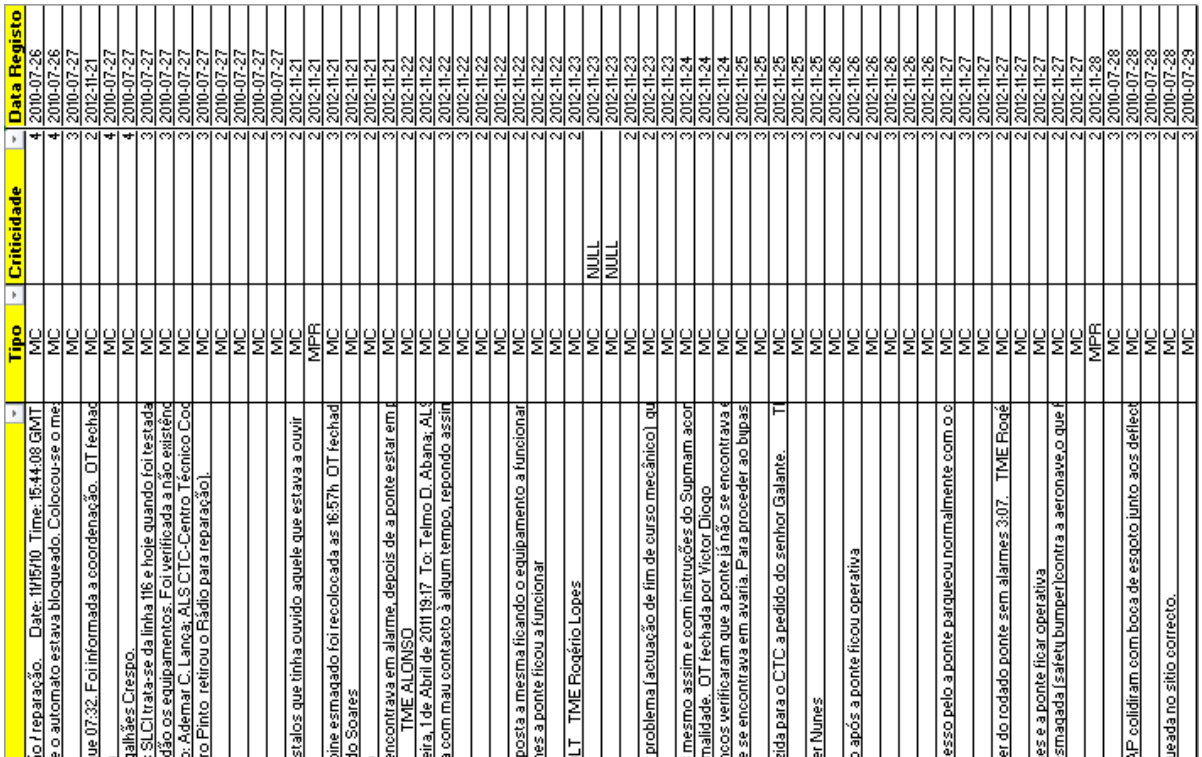
27.06.2017 Salida dinámica de lista 1

S	Cl.	Inic.prog.	Fin progr.	Orden	Ubicación técnica	Equipo	Denominación de la ubicación técnica	Texto breve	P	Mant.Prv	P	PtoTrDRes	TotGenPlan	TotGenReal	Status del sistema	
	COR1			10000001529	P794-PM1-P40-00900	45200001000001	ASTILLADORA	sus. placa desgaste cuchilla			2	LINMEC	0.00	2,097.50		
	COR1			10000001615	P794-PM1-P40-00900		ASTILLADORA	sustitución correas astilladora			2	LINMEC	0.00	9,861.80		
	COR1			10000001701	P794-PM1-P40-00900		ASTILLADORA	"sus rodillo pinchos astilladora"			2	LINMEC	0.00	2,248.00		
	COR1			10000001826	P794-PM1-P40-00900		ASTILLADORA	limpieza mandibula astilladora			2	LINMEC	0.00	502.52		
	COR1			10000002543	P794-PM1-P40-00900		ASTILLADORA	sus. correas astilladora			2	LINMEC	0.00	10,141.57		
	COR1			10000002584	P794-PM1-P40-00900		ASTILLADORA	sus. rodillo pinchos alimet astilladora			5	LINMEC	0.00	5,252.23		
	COR1			10000002618	P794-PM1-P40-00900		ASTILLADORA	SUS. PEINE ENTRADA ASTILLADORA			5	LINMEC	0.00	2,208.35		
	COR1			10000002618	P794-PM1-P40-00900		ASTILLADORA	SUS. CONTRACUCHILLA ASTILLADORA			5	LINMEC	0.00	879.48		
	COR1			10000004197	P794-PM1-P40-00900		ASTILLADORA	REPOSICION RELE SEGURIDAD ASTILLADORA			7	LINELEC	0.00	209.80		
	COR1			10000004236	P794-PM1-P40-00900		ASTILLADORA	sus. correas x 13 usadas			2	LINMEC	0.00	153.61		
	PRV2	10.04.2012		10000021359	P794-PM1-P40-00900		ASTILLADORA	Sus fresa ejes aliment astilladora ***			6	LINMEC	0.00	1,757.60	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	14.06.2012		10000022684	P794-PM1-P40-00900		ASTILLADORA	Rep trans mandibula sup astilladora			1	LINMEC	0.00	4,906.49	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	15.06.2012		10000022709	P794-PM1-P40-00900		ASTILLADORA	Sus pinchos rodilla sup/inf ultimo			1	LINMEC	0.00	4,063.72	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	03.07.2012		10000023860	P794-PM1-P40-00900		ASTILLADORA	Rep eje rodillo aceleración entra astill			1	LINMEC	0.00	1,077.72	CTEC INPR KKMP MOVM NLIQ PREC	
	COR1	20.08.2012		10000023878	P794-PM1-P40-00900	45200001000001	ASTILLADORA	REPARACIÓN CABLE 6kV ALIMENTACIÓN ASTILL			1	LINMEC	0.00	2,782.40	CTEC KKMP MOVM NLIQ PREC	
	PRV1	24.09.2012		10000024532	P794-PM1-P40-00900	45200001000001	ASTILLADORA	LEVANTAR CAPOTA DE ASTILLADORA PARA LIMP			1	LINMEC	0.00	695.07	CTEC KKMP MOVM NLIQ PREC	
	PRV2	08.03.2013		10000027447	P794-PM1-P40-00900		ASTILLADORA	SUST.PIÑON DE ARRASTRE DE MOTOR Y REDUCT			6	LINMEC	0.00	46.15	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV1	23.04.2013		10000028321	P794-PM1-P40-00900		ASTILLADORA	VACIADO FOSO DE LA ASTILLADORA			6	PLANIF	0.00	440.00	CTEC KKMP MOVM NLIQ PREC	
	PRV2	28.05.2013		10000028928	P794-PM1-P40-00900		ASTILLADORA	SUST. TAMIZ DE ASTILLADORA			4	LINMEC	0.00	2,058.57	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	23.07.2013		10000029869	P794-PM1-P40-00900		ASTILLADORA	Sus. protecciones amortiguadores soporta			6	LINMEC	0.00	196.20	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV1	18.09.2013		10000030441	P794-PM1-P40-00900		ASTILLADORA	LIMPIEZA FOSO ASTILLADORA			6	PLANIF	288.20	288.20	CTEC KKMP MOVM NLIQ PREC	
	PRV2	27.09.2013		10000030887	P794-PM1-P40-00900		ASTILLADORA	SUST.DIENTES RODILLOS PIÑA ASTILLADORA.(			6	LINMEC	0.00	5,297.81	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	30.09.2013		10000030996	P794-PM1-P40-00900		ASTILLADORA	SUST.TRANSMISION RODILLO PIÑA MANOBIULA (			6	LINMEC	0.00	371.02	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	30.10.2013		10000031430	P794-PM1-P40-00900		ASTILLADORA	FABRICACION Y MONTAJE MUNEQUILLA RODILLO			6	LINMEC	175.00	209.65	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	17.12.2013		10000032243	P794-PM1-P40-00900		ASTILLADORA	SUST.CADENA TRANSMISION RODILLOS PIÑA MA			6	LINMEC	0.00	174.44	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	13.02.2014		10000033103	P794-PM1-P40-00900		ASTILLADORA	SUST.ZAPATAS CUCHILLAS ROTOR ASTILLADORA			6	LINMEC	0.00	3,176.40	LIB. INPR KKMP MOVM NLIQ PREC	
	PRV2	11.06.2014		10000033103	P794-PM1-P40-00900		ASTILLADORA	SUST.CONTRACUCHILLAS ASTILLADORA. (MOTOR			6	LINMEC	860.00	2,609.82	LIB. INPR KKMP MOVM NLIQ PREC	
	COR1	26.02.2014		10000033200	P794-PM1-P40-00900		ASTILLADORA	REPARACION DE PISTON EXTRACTOR CONTRACUC			6	LINMEC	795.00	934.20	LIB. INPR KKMP MOVM NLIQ PREC	
	PRV2	09.05.2014		10000033781	P794-PM1-P40-00900		ASTILLADORA	SUST.PEINE ASTILLADORA. (MOTOR 01 150)]			6	LINMEC	926.42	2,674.52	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	25.04.2014		10000034159	P794-PM1-P40-00900		ASTILLADORA	LIMPIEZA MANOBIULA RODILLOS PIÑA ASTILLA			6	LINMEC	0.00	139.20	LIB. INPR KKMP MOVM NLIQ PREC	
	PRV1	14.05.2014		10000034405	P794-PM1-P40-00900	45200001000001	ASTILLADORA	SUSTITUCION ZAPATA ROTOR ASTILLADORA			3	LINMEC	0.00	0.00	CTEC INPR KKMP NLIQ PREC	
	PRV2	26.05.2014		10000034564	P794-PM1-P40-00900		ASTILLADORA	SUST.MOTORREDUCTOR RODILLO ACELERADOR DE			6	LINMEC	0.00	518.93	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	30.05.2014		10000034630	P794-PM1-P40-00900		ASTILLADORA	EXTRACCION CUCHILLA ROTOR ASTILLADORA, S			6	LINMEC	0.00	2,294.86	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	20.06.2014		10000034665	P794-PM1-P40-00900		ASTILLADORA	REPARACION PISTON EXTRACTOR DE CONTRACUC			6	LINMEC	325.00	879.38	CTEC INPR KKMP MOVM NLIQ PREC	
	COR1	02.07.2014		10000034720	P794-PM1-P40-00900		ASTILLADORA	REPARACION BOTONERA AEREA POLIPASTO DE C			7	LINELEC	0.00	730.76	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	12.08.2014		10000034902	P794-PM1-P40-00900		ASTILLADORA	SUST.EJE RODILLO ACELERADOR DE TACOS ENT			6	LINMEC	165.00	304.20	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	01.07.2014		10000035059	P794-PM1-P40-00900		ASTILLADORA	REPARAR MACACO DE LA ASTILLADORA			6	LINMEC	125.90	125.90	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	22.07.2014		10000035336	P794-PM1-P40-00900		ASTILLADORA	REPARACION CAPOTA DE ASTILLADORA (MOTOR			6	LINMEC	125.90	383.65	LIB. INPR KKMP MOVM NLIQ PREC	
	PRV2	04.11.2014		10000036657	P794-PM1-P40-00900		ASTILLADORA	SUST.MUÑEQUILLA RODILLO ACELERADOR DE TA			3	LINMEC	165.00	165.00	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	21.11.2014		10000036706	P794-PM1-P40-00900		ASTILLADORA	SUSPENSION DE CAPOTA PARA LIMPIEZA DE MA			6	LINMEC	594.04	798.69	CTEC INPR KKMP MOVM NLIQ PREC	
	COR1	19.11.2014		10000036891	P794-PM1-P40-00900		ASTILLADORA	REPONER CABLES ZONA ENTRADA ASTILLADORA			7	LINELEC	0.00	114.80	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	24.11.2014		10000036951	P794-PM1-P40-00900		ASTILLADORA	SUST.PINCHOS PENULTIMO RODILLO INFERIOR			6	LINMEC	0.00	2,482.65	LIB. INPR KKMP MOVM NLIQ PREC	
	PRV2	04.12.2014		10000037102	P794-PM1-P40-00900		ASTILLADORA	SUST. LATIGUILLO MACACOS ASTILLADORA (NO			6	LINMEC	0.00	99.53	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	05.12.2014		10000037113	P794-PM1-P40-00900		ASTILLADORA	SUST.CORREAS ROTOR ASTILLADORA (MOTOR 01			6	LINMEC	0.00	3,390.10	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	16.12.2014		10000037267	P794-PM1-P40-00900		ASTILLADORA	RETENSADO CORREAS TRANSMISION ASTILLADOR			6	LINMEC	0.00	44.00	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	11.02.2015		10000037955	P794-PM1-P40-00900		ASTILLADORA	SUST. MUÑEQUILLA RODILLO ACELERADOR DE T			6	LINMEC	115.00	115.00	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	24.02.2015		10000038128	P794-PM1-P40-00900		ASTILLADORA	Reparación perno sujección tope de segur			6	LINMEC	75.00	214.20	CTEC INPR KKMP MOVM NLIQ PREC	
	COR1	12.03.2015		10000038321	P794-PM1-P40-00900		ASTILLADORA	REPARACION DEL 5º EJE INFERIOR DE LA AST			7	LINMEC	897.84	3,877.74	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	20.03.2015		10000038465	P794-PM1-P40-00900		ASTILLADORA	SUST.CADENA DE TRANSMISION RODILLOS PIÑA			6	LINMEC	0.00	0.00	CTEC INPR KKMP NLIQ PREC	
	PRV2	25.03.2015		10000038553	P794-PM1-P40-00900		ASTILLADORA	RETENSADO CORREAS ASTILLADORA,ELIMINACIO			6	LINMEC	0.00	0.00	CTEC INPR KKMP NLIQ PREC	
	COR1	01.04.2015		10000038580	P794-PM1-P40-00900	45200001000001	ASTILLADORA	MONTAR PROTECCION ASTILLADORA PARA CABLE			6	LINELEC	37.00	39.20	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	31.03.2015		10000038618	P794-PM1-P40-00900		ASTILLADORA	SUSTITUIR 2 PIÑONES MANOBIULA SUPERIOR D			4	LINMEC	975.00	2,261.24	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	09.04.2015		10000038745	P794-PM1-P40-00900		ASTILLADORA	SUSTITUCION DE CORREAS DE TRANSMISION DE			6	LINMEC	0.00	3,333.10	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	10.04.2015		10000038767	P794-PM1-P40-00900		ASTILLADORA	RETENSADO DE CORREAS DE TRANSMISION DE L			6	LINMEC	0.00	0.00	CTEC INPR KKMP NLIQ PREC	
	PRV2	16.08.2016		10000038981	P794-PM1-P40-00900		ASTILLADORA	REPARACION DE PISTONES ELEVACION DE MAND			5	LINMEC	0.00	0.00	CTEC INPR KKMP NLIQ PREC	
	PRV2	24.04.2015		10000038981	P794-PM1-P40-00900		ASTILLADORA	SUSTITUCION DE RETEN EN REDUCTOR MANOBIU			4	LINMEC	0.00	37.80	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	11.05.2015		10000039205	P794-PM1-P40-00900		ASTILLADORA	SUSTITUIR PIEZA DE DESGASTE AMARRE CUCHI			4	LINMEC	0.00	628.65	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	12.05.2015		10000039205	P794-PM1-P40-00900		ASTILLADORA	SUSTITUIR SILEMBLOCK REDUCTORES DE ACCIO			6	LINMEC	0.00	244.68	CTEC INPR KKMP MOVM NLIQ PREC	
	COR1	13.05.2015		10000039234	P794-PM1-P40-00900	45200001000001	ASTILLADORA	SUSTITUIR CABEZA FINAL DE CARRERA CAPOTA			7	LINELEC	0.00	10.43	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	08.06.2015		10000039626	P794-PM1-P40-00900		ASTILLADORA	SUSTITUCION CORREAS DE TRANSMISION DE LA A			4	LINMEC	0.00	3,819.34	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	07.07.2015		10000039845	P794-PM1-P40-00900		ASTILLADORA	SUSTITUCION DE TAMIZ DE ASTILLADORA			4	LINMEC	198.00	949.06	CTEC INPR KKMP MOVM NLIQ PREC	
	COR1	14.07.2015		10000040194	P794-PM1-P40-00900		ASTILLADORA	REPONER INTERRUPTOR CAMPO POLIPASTO ASTI			7	LINELEC	0.00	46.87	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	16.07.2015		10000040224	P794-PM1-P40-00900		ASTILLADORA	SUST. CADENA DE TRANSMISION RODILLOS PIÑ			6	LINMEC	0.00	102.08	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	27.07.2015		10000040228	P794-PM1-P40-00900		ASTILLADORA	SUST. EJE ULTIMO RODILLO PIÑA MANOBIULA (			6	LINMEC	153.70	3,717.76	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	30.09.2015		10000041263	P794-PM1-P40-00900		ASTILLADORA	SUST. PLACAS DE DESGASTE CUCHILLAS ASTIL			4	LINMEC	0.00	1,518.90	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	21.10.2015		10000041263	P794-PM1-P40-00900		ASTILLADORA	INSPECCION TRANSMISION MANOBIULA SUPERIO			4	LINMEC	0.00	278.40	CTEC INPR KKMP MOVM NLIQ PREC	
	COR1	09.10.2015		10000041386	P794-PM1-P40-00900		ASTILLADORA	SUSTITUCION CABEZA FINAL CARRERA CAPOTA			6	LINELEC	0.00	10.53	CTEC INPR KKMP MOVM NLIQ PREC	
	PRV2	03.11.2015		10000041655	P794-PM1-P40-00900		ASTILLADORA	SUSTITUIR SPIKE´S EN RODILLO INFERIOR Nº			6	LINMEC	25.16	2,154.36	CTEC INPR KKMP MOVM NLIQ PREC	

2º Dataset

	OT	Descrição OT	Descrição Log
2	ALS1000011	Ponte 144 em avaria e com falha de autolevel	NULL
3	ALS1000048	varia da manga 146	saltu edge actuado, recolocada a manga ficando operativa
4	ALS1000053	Avaria manga 143	puxada a manga atrás e retirado o alarme ficando operativa
5	ALS1000715	Reporte de ponte 146.2 fora de parque	Reporte de ponte 146.2 fora de parque
6	ALS1000100	Avaria Ponte Telescópica FP-.TEK 146.2	o operador tinha dificuldade em manobrar a manga
7	ALS1000100	Avaria Ponte Telescópica FP-.TEK 146.2	o operador tinha dificuldade em manobrar a manga
8	ALS1000111	ALARME GERAL NA MANGA 145	retirado alarme ficando operativa
9	ALS1000113	ALARME GERAL NA PONTE 142	retirado alarme ficando operativa
10	ALS1000113	ALARME GERAL NA PONTE 142	retirado alarme ficando operativa
11	ALS1000135	Manga 146 com anomalia	corrigida anomalia
12	ALS1000142	Camara da manga virada ao contrario	Ver Comentário acima.
13	ALS1000142	Camara da manga virada ao contrario	Relatório
14	ALS1000142	Camara da manga virada ao contrario	Relatório
15	ALS1000146	Manga 116 - Falha de Auto Level	CORRIGIDA ANOMALIA
16	ALS1004456	Ponte 146.2 ao finalizar acostamento em manual ao rodar cabine para Esq faz ruido (estalos)	NULL
17	ALS1010063	PMR SPLITS MANGAS TELESCÓPICAS	NULL
18	ALS1011688	Ponte 122 em avaria	Ponte 122 em avaria
19	ALS1011689	Ponte 116 em alarme geral	Ponte 116 em alarme geral
20	ALS1011696	Ponte 122 em alarme geral	Ponte 122 em alarme geral
21	ALS1011704	Ponte 126 em alarme geral /falha de autonivelamento	Ponte 126 em alarme geral /falha de autonivelamento
22	ALS1011717	Ponte 142- am alarme geral /falha de auto nivelamento	Limpeza de alarmes e reposição do funcionamento
23	ALS1011717	Ponte 142- am alarme geral /falha de auto nivelamento	Limpeza de alarmes e reposição do funcionamento
24	ALS1011722	Parte da iluminação do tunel da ponte 123 não	Verificação do contactor da iluminação da manga 123
25	ALS1011722	Parte da iluminação do tunel da ponte 123 não	Verificação do contactor da iluminação da manga 123
26	ALS1011732	Parte da iluminação do tunel da ponte 123 não	Parte da iluminação do tunel da ponte 123 apagada
27	ALS1011745	Ponte 142 em alarme geral	Ponte 142 em alarme geral
28	ALS1011745	Ponte 142 em alarme geral	Ponte 142 em alarme geral
29	ALS1011777	Ponte 142 em alarme geral	NULL
30	ALS1011820	ponte 126 em alarme geral	ponte 126 em alarme geral
31	ALS1011830	Ponte 116 em alarme geral	Ponte 116 em alarme geral
32	ALS1011877	FALHA DE AUTONIVELAMENTO PONTE 122	NULL
33	ALS1011891	FACILITADO ACOMPANHAMENTO A VISUALIZAÇÃO DAS PONTES 115 E 146	NULL
34	ALS1011934	ALARME GERAL NA PONTE 122	ALARME GERAL NA PONTE 122
35	ALS1011934	ALARME GERAL NA PONTE 122	ALARME GERAL NA PONTE 122
36	ALS1011907	ALARME GERAL NA PONTE 122	ALARME GERAL NA PONTE 122
37	ALS1011938	Ponte 122 em falha de auto-level	NULL
38	ALS1011958	Ponte 114 em alarme geral	Ponte 114 em alarme geral
39	ALS1011976	Ponte 122 em alarme geral	Ponte 122 em alarme geral
40	ALS1011981	FALHA DE ALIMENTAÇÃO PONTE 142	Falha de energia provocada pela remoção da UPS em avaria
41	ALS1011981	FALHA DE ALIMENTAÇÃO PONTE 142	Falha de energia provocada pela remoção da UPS em avaria
42	ALS1011983	ANOMALIA NA UPS DA PONTE 142	Remoção da UPS da ponte 142
43	ALS1011983	ANOMALIA NA UPS DA PONTE 142	Remoção da UPS da ponte 142
44	ALS1012021	Ponte 115 em alarme geral /falha de auto nivelamento	Ponte 115 em alarme geral /falha de auto nivelamento
45	ALS1012077	Ponte 123 AC a funcionar mas a temperatura esta baixa	NULL
46	ALS1012110	Ponte 144 em alarme geral	NULL
47	ALS1012112	Ponte 142 em alarme geral	NULL
48	ALS1012128	Ponte 122 em alarme geral	Ponte 122 em alarme geral
49	ALS1012130	pontes 146.1 e 2 em falha	pontes 146.1 e 2 em falha
50	ALS1012135	Ponte 122 em alarme geral	Ponte 122 em alarme geral
51	ALS1012206	Ponte 114 em Alarme geral /falha de auto nivelamento	NULL
52	ALS1012209	Ponte 142 em alarme geral /falha de auto nivelamento	Ponte 142 em alarme geral /falha de auto nivelamento
53	ALS1012210	Ponte 146 - desencostou em automatico e só anda em manual muito lentamente	Ponte 146 - desencostou em automatico e só anda em manual muito lentamente
54	ALS1012211	Ponte 142 em alarme geral sem aeronave	Ponte 142 em alarme geral sem aeronave
55	ALS1012249	Calor na manga 123	NULL
56	ALS1012268	Ponte 122 em alarme geral	NULL
57	ALS1012272	Ponte 122 em alarme geral no voo Air France 1624	NULL
58	ALS1012281	Muito frio na ponte 146.1	Muito frio na ponte 146.1
59	ALS1013175	PME A PTE FP-TEK 116	NULL
60	ALS1016681	ANOMALIA NA MANGA 146.2	CORRIGIDA ANOMALIA
61	ALS1016691	ANOMALIA NA MANGA 146.2	corrigida anomalia auto-level
62	ALS1016691	ANOMALIA NA MANGA 146.2	corrigida anomalia auto-level
63	ALS1017718	Manga 145 não tem selecção de A319	Manga 145 não tem selecção de A319
64	ALS1017730	ANOMALIA NA MANGA 143	houve uma má acostagem da parte so operador fazendo com que a borracha fosse actuada.

2º Dataset (cont.)



Descrição Longa Log	Tipo	Criticidade	Data Registo
Foi verificada a UPS e no local está sem alarmes.. Solicita-se verificação / reparação.... Date: 11/15/10 Time: 15:44:08 GMT	MC	4	2010-07-26
Depois de verificações a varios sistemas de sequrança concluiu-se que o automato estava bloqueado. Colocou-se o me	MC	4	2010-07-26
NULL	MC	3	2010-07-27
Foram verificadas as duas pontes 146.1 e146.2 estando as duas em parque 07:32. Foi informada a coordenação.. OT fechad	MC	2	2012-11-21
foi dada alguma informação de como proceder ficando bem encostada	MC	4	2010-07-27
Conforme falado junto envio fotografia.. Cumprimentos, Tomaz Magalhães Crespo.	MC	4	2010-07-27
Esta HOT LINE é testada 2 vezes por dia pelo pessoal do SLC1 Para os SLC1 trata-se da linha 116 e hoje quando foi testada	MC	3	2010-07-27
Foi desligada a alimentação e retirados os 3 conjuntos de fusíveis que dão os equipamentos. Foi verificada a não existênc	MC	3	2010-07-27
From: Paulo A. Fradique Sent: domingo, 15 de Agosto de 2010 14:40 To: Ademar C. Lança; ALS CTC-Centro Técnico Coc	MC	3	2010-07-27
Solicitou-se verificação / correcção ao TEA de serviço.. [o colega Pedro Pinto retirou o Radio para reparação].	MC	3	2010-07-27
NULL	MC	2	2010-07-27
Foi ligada a UTA 3	MC	2	2010-07-27
Quando la chegamos verificamos que estava tudo operacional.	MC	2	2010-07-27
Solicita-se correcção/verificação da porta 5.14.204	MC	3	2010-07-27
Foram feitos testes ao foi metálico ao que a operadora adicionou os estalos que tinha ouvido aquele que estava a ouvir	MC	2	2012-11-21
NULL	MPR	2	2012-11-21
Foi necessario puxar a ponte a trás porque tinha o safetu bumper de cabine esmagado foi recolocada as 16:57h. OT fechad	MC	3	2012-11-21
Foi aceite e reposto o alarme ficando operativa. OT fechada por Eduardo Soares	MC	3	2012-11-21
Foi aceite os alarmes e reposta às 19:52h. OT fechada por Victor Diogo	MC	2	2012-11-21
há chegada ao local i o operador estava a retirara ponte mas ainda se encontrava em alarme, depois de a ponte estar em p	MC	3	2012-11-21
Foram limpos os alarmes e reposta a sua funcionalidade na Ponte 142. TME ALONSO	MC	2	2012-11-22
From: Ocorrencias@ana.pt [mailto:Ocorrencias@ana.pt] Sent: sexta-feira, 1 de Abril de 2011 19:17 To: Telmo D. Abana; ALS	MC	2	2012-11-22
Foi verificado o contactor da iluminação da manga 123, que se encontra com mau contacto à algum tempo, repondo assin	MC	3	2012-11-22
Solicita-se verificação / correcção.	MC	3	2012-11-22
Foram substituidas 22 lâmpadas.	MC	2	2012-11-22
aceite os alarme s as 7:38, a ponte encontrava-se em parque	MC	2	2012-11-22
O elevador tinha a botoneira de paragem de emergência actuada. Foi reposta a mesma ficando o equipamento a funcionar	MC	2	2012-11-22
foram verificadas as fichas dos Bumperes das rodas,retirados os alarmes a ponte ficou a funcionar	MC	2	2012-11-22
resolvido pelo operador.. Ot fechada por Valter Nunes	MC	2	2012-11-22
Ponte 116 em alarme geral.Alarme retirado pelo Tecnico Alonso às 7:59LT TME Rogério Lopes	MC	2	2012-11-23
NULL	MC	NULL	2012-11-23
NULL	MC	NULL	2012-11-23
FICOU OPERATIVO ÀS 15H39 SEM DL-87.	MC	2	2012-11-23
Hoje pelas 19:58 na chegada do voo EZY 7996 à manga 122 ocorreu um problema [actuação de fim de curso mecânico] qu	MC	2	2012-11-23
FICOU OPERATIVO ÀS 19H40 SEM DL-87. Alexandre Filipe	MC	3	2012-11-23
Ao chegar ao local já o operador tinha conseguido retirar o alarme mas mesmo assim e com instruções do Supmam acor	MC	3	2012-11-24
Tinha sido atuada a paragem de emergência. Foi aceite e reposto a normalidade.. OT fechada por Victor Diogo	MC	3	2012-11-24
Alarme verificado na supervisão às 23:14LT. À chegada ao local os técnicos verificaram que a ponte já não se encontrava e	MC	2	2012-11-24
A falha de energia foi provocada pela remoção da UPS da ponte 142 que se encontrava em avaria. Para proceder ao bypas	MC	3	2012-11-25
Rearmados os circuitos 16 e 7 do 1QF F2.3.7	MC	3	2012-11-25
Foi removida a UPS da ponte 142 e colocada em Bypass. A UPS foi trazida para o CTC a pedido do senhor Galante. TI	MC	3	2012-11-25
Foi rearmada 02-04-2011 11:10	MC	3	2012-11-25
Corrigido e aceitação de alarmes Ponte OK 19:18 Ot fechada por Valter Nunes	MC	3	2012-11-25
foi aumentada a temperatura no quente	MC	2	2012-11-26
Verificação e limpeza das fichas e em sequida retirada de alarmes e logo após a ponte ficou operativa	MC	2	2012-11-26
Reposição de alarmes	MC	2	2012-11-26
Foi retirado o alarme.. Ot fechada por Valter Nunes	MC	3	2012-11-26
Foi retirada a falha PBB Link.. Ot fechada por Valter Nunes	MC	3	2012-11-26
sem alarme 18:12 Alexandre Filipe	MC	3	2012-11-26
À chegada ao local a ponte encontrava-se com alarme de cartao se acesso pelo a ponte parqueou normalmente com o c	MC	2	2012-11-27
Corrigida a manobra/ aceitação ponte sem alarmes 23:41	MC	3	2012-11-27
Verificada ponte em parque sem qualquer anomalia 23:45	MC	3	2012-11-27
Ponte 142 em alarme geral sem aeronave.Verificadas fichas dos Bumper do rodado ponte sem alarmes 3:07.. TME Rogê	MC	2	2012-11-27
baixar temperatura no A.C. da manga	MC	2	2012-11-27
o TME teve que esperar pelo fim do embarque para poder tirar os alarmes e a ponte ficar operativa	MC	2	2012-11-27
Ponte 122 em alarme geral no voo Air France 1624.Ponte com cabine esmagada [safetu bumper]contra a aeronave,o que f	MC	2	2012-11-27
corrigidos os setpoints pelos tecnicos do CMT ok às 17:50LT	MC	2	2012-11-27
Efectuado	MPR	2	2012-11-28
NULL	MC	3	2010-07-28
Aeroporto LIS-SOA. Data 2012-02-27 08:00:00 Descrição Dolleys da TAP colidiram com boca de esgoto junto aos deflect	MC	3	2010-07-28
Rearmado o ID 2 que se encontrava disparado	MC	3	2010-07-28
Não foi possivel fazer o "teaching" porque a aeronave não estava parqueada no sitio correcto.	MC	2	2010-07-28
Solicita-se reparação.	MC	3	2010-07-29

3º Dataset (tabela de alarmes)

	datetime	machineID	errorID
0	2015-01-06 03:00:00	1	error3
1	2015-02-03 06:00:00	1	error4
2	2015-02-21 11:00:00	1	error1
3	2015-02-21 16:00:00	1	error2
4	2015-03-20 06:00:00	1	error1
5	2015-04-04 06:00:00	1	error5
6	2015-05-04 06:00:00	1	error4
7	2015-05-19 06:00:00	1	error2
8	2015-05-19 06:00:00	1	error3
9	2015-06-03 06:00:00	1	error5
10	2015-06-18 06:00:00	1	error2

3º Dataset (tabela de falhas)

	datetime	machineID	failure
0	2015-02-04 06:00:00	1	comp3
1	2015-03-21 06:00:00	1	comp1
2	2015-04-05 06:00:00	1	comp4
3	2015-05-05 06:00:00	1	comp3
4	2015-05-20 06:00:00	1	comp2
5	2015-06-04 06:00:00	1	comp4
6	2015-06-19 06:00:00	1	comp2
7	2015-08-03 06:00:00	1	comp3
8	2015-08-03 06:00:00	1	comp4
9	2015-11-01 06:00:00	1	comp4
10	2015-11-16 06:00:00	1	comp1

3º Dataset (Tabela de equipamentos)

	machineID	model	age
0	1	model2	18
1	2	model4	7
2	3	model3	8
3	4	model3	7
4	5	model2	2
5	6	model3	7
6	7	model4	20
7	8	model3	16
8	9	model1	7
9	10	model1	10
10	11	model4	6

3º Dataset (Tabela de manutenção)

	datetime	machineID	comp	IF_FAIL
0	2014-07-01 06:00:00	1	comp4	0
1	2014-09-14 06:00:00	1	comp1	0
2	2014-09-14 06:00:00	1	comp2	0
3	2014-11-13 06:00:00	1	comp3	0
4	2015-01-05 06:00:00	1	comp1	0
5	2015-01-20 06:00:00	1	comp1	0
6	2015-02-04 06:00:00	1	comp3	1
7	2015-02-19 06:00:00	1	comp3	0
8	2015-03-06 06:00:00	1	comp3	0
9	2015-03-21 06:00:00	1	comp1	1
10	2015-04-05 06:00:00	1	comp1	0

3º Dataset (Tabela de sensorização)

	datetime	machineID	volt	rotate	pressure	vibration
0	2015-01-01 06:00:00	1	151.919999	530.813578	101.788175	49.604013
1	2015-01-01 07:00:00	1	174.522001	535.523532	113.256009	41.515905
2	2015-01-01 08:00:00	1	146.912822	456.080746	107.786965	42.099694
3	2015-01-01 09:00:00	1	179.530561	503.469990	108.283817	37.847727
4	2015-01-01 10:00:00	1	180.544277	371.600611	107.553307	41.467880
5	2015-01-01 11:00:00	1	141.411757	530.857266	87.614001	44.985846
6	2015-01-01 12:00:00	1	184.083822	450.227529	87.697380	30.831263
7	2015-01-01 13:00:00	1	166.632618	486.466838	108.067734	50.380054
8	2015-01-01 14:00:00	1	159.892748	488.968697	102.131884	43.661297
9	2015-01-01 15:00:00	1	176.686812	508.202759	90.951189	43.039696
10	2015-01-01 16:00:00	1	189.209257	489.650433	105.406282	41.822968

ANEXO B:Base de dados de OTs do projeto UNO